1

THE CAPITALIST REVOLUTION

© Kevin Poh, Flickr

HOW CAPITALISM REVOLUTIONISED THE WAY WE LIVE, AND HOW ECONOMICS ATTEMPTS TO UNDERSTAND THIS AND OTHER ECONOMIC SYSTEMS

- There have been dramatic changes in living standards in different countries in the last 1,000 years
- In many countries these living standards began to rise rapidly at the time of the capitalist revolution
- Advances in technology and a distinctive economic system contributed to this revolution
- Economics is the study of how people interact with each other, and with the natural environment, in producing their livelihoods
- Capitalism is an economic system in which private property, markets and firms play a major role
- The rise in living standards has been accompanied by changes in population and the way people live, by environmental impacts, and by changes in inequality between countries and within countries
- There is great variation across countries in their success in raising incomes, and in the degree of inequality in living standards within them

See www.core-econ.org for the full interactive version of *The Economy* by The CORE Project. Guide yourself through key concepts with clickable figures, test your understanding with multiple choice questions, look up key terms in the glossary, read full mathematical derivations in the Leibniz supplements, watch economists explain their work in Economists in Action – and much more.

Institute for New Economic Thinking
UCL SciencesPo Azim Premji University
FRIENDS PROVIDENT Foundation

In the 14th century the Moroccan scholar Ib'n Battuta (see box) described Bengal in India as "a country of great extent, and one in which rice is extremely abundant. Indeed, I have seen no region of the earth in which provisions are so plentiful." And he had seen much of the world, having travelled to China, west Africa, the Middle East and Europe. Three centuries later, the same sentiment was expressed by the 17th century French diamond merchant Jean Baptiste Tavernier who wrote of the country:

"Even in the smallest villages, rice, flour, butter, milk, beans and other vegetables, sugar and sweetmeats, dry and liquid, can be procured in abundance…"
– Jean Baptiste Tavernier, *Travels in India* (1676)

At the time of Ib'n Battuta's travels India was not richer than the other parts of the world. But India was not much poorer, either. An observer at the time would have noticed that people, on average, were better off in Italy, China and England than in Japan or India. But the vast differences between the rich and the poor, which the traveller would have noted wherever he went, were much more striking than these differences across regions. Rich and poor would often have different titles: in some places they would be feudal lords and serfs, in others royalty and their subjects, slave owners and slaves, or merchants and the sailors who transported their goods. Then—as now—your prospects depended on where your parents were on the economic ladder and whether you were male or female. The difference in the 14th century, compared with today, was that then it mattered much less in which part of the world you were born.

Fast forward to today. The people of India are far better off than they were seven centuries ago if we think about their access to food, medical care, shelter and the necessities of life; but by world standards today most are poor.

IB'N BATTUTA

IB'N BATTUTA

Ib'n Battuta (1304-1368) was a Moroccan traveller and merchant whose travels were published in his book *Rihla* (The Journey). His travels, lasting 30 years, took him across north and west Africa, eastern Europe, the Middle East, south and central Asia and China. He travelled more than 70,000 miles (113,000km); much further than the distance covered by his better-known contemporary, Marco Polo (1254-1324).

Figure 1.1a tells some of the story (you can follow links from the figure to the sources of the data). The height of each line is an estimate of average living standards—using a measure called *gross domestic product per capita*, which we will explain in the next section—at the date on the horizontal axis.

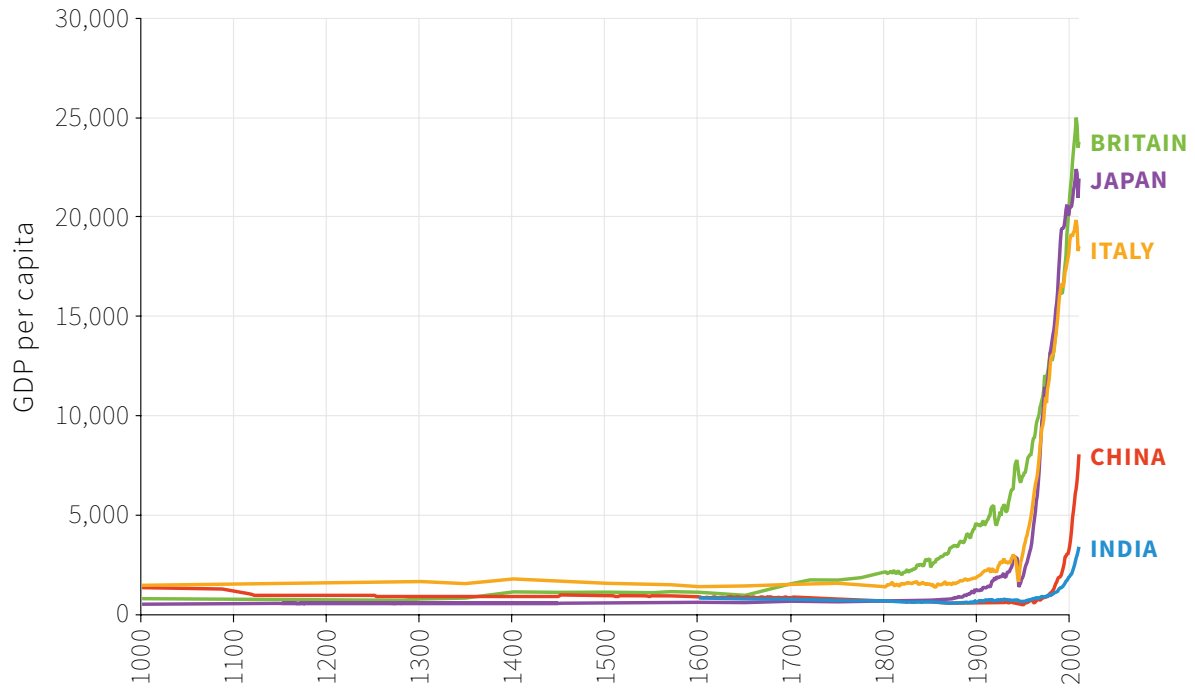


Figure 1.1a *History's hockey stick: Gross domestic product per capita in five countries (1000-2013).*

Source: *Bolt, Jutta, and Jan Juiten van Zanden. 2013. 'The First Update of the Maddison Project Re-Estimating Growth Before 1820.' Maddison-Project Working Paper WP-4. Broadberry, Stephen. 2013. 'Accounting for the Great Divergence.' London School of Economics and Political Science. November 1.*

On average people in the UK are six times better off than in India by this measure. Japanese people are as rich as the British, just as they were in the 14th century, but now Americans are even better off than the Japanese, and Norwegians are better off still.

We can draw the graph in Figure 1.1a because of the work of Angus Maddison who dedicated his working life to finding the scarce data to make useful comparisons of how people lived across more than 1,000 years (his work is continuing in the Maddison Project). In this course you will see that data like this about regions of the world, and the people in it, is the starting point of all economics: in this video, Nobel laureate James Heckman and Thomas Piketty explain how collecting data has been fundamental to their work on inequality and the policies to reduce it. We will study their work in Unit 19.

So 1,000 years ago the world was flat, economically speaking. There were differences in income between the regions of the world; but as you can see from Figure 1.1a, the differences were small compared to what was to follow.

1.1 HISTORY'S HOCKEY STICK: GROWTH IN INCOME

A different way of looking at the same data in Figure 1.1a is to use a scale that shows GDP per capita doubling as we move up the vertical axis (from $250 per capita per year to $500, then to $1,000, and so on). This is called a *ratio scale* and is shown in Figure 1.1b. The ordinary scale is useful for comparing the levels of GDP per capita across countries, but the ratio scale is best for comparing growth rates across countries.

By the *growth rate* of GDP or of any other quantity like population, we mean the rate of change:

$$\textit{growth rate} = \frac{\textit{change in GDP}}{\textit{original level of GDP}}$$

If the level of GDP per capita in the year 2000 is $21,046, as it was in Britain in the data shown in Figure 1.1a, and $21,567 in 2001, then we can calculate the growth rate:

$$\begin{aligned}\textit{growth rate} &= \frac{\textit{change in GDP}}{\textit{original level of GDP}} \\ &= \frac{y_{2001} - y_{2000}}{y_{2000}} \\ &= \frac{21{,}157 - 21{,}406}{21{,}406} \\ &= 0.025 \\ &= 2.5\%\end{aligned}$$

Whether we want to compare levels or growth rates depends on the question we are asking. Figure 1.1a makes it easy to compare the levels of GDP per capita across countries, and at different times in history. Figure 1.1b uses a ratio scale, which makes it possible to compare growth rates across countries and at different periods. When a ratio scale is used, a series that grows at a constant rate looks like a straight line. This is because the percentage (or proportional growth rate) is constant. A steeper line in the ratio scale chart means a faster growth rate.

To see this, think of a growth rate of 100%: that means the level doubles. In Figure 1.1b, with the ratio scale, you can check that if GDP per capita doubled over 100 years from a level of $500 to $1,000, the line would have the same slope as a doubling from $2,000 to $4,000 dollars, or from $16,000 to $32,000 over 100 years. If, instead of doubling, the level quadrupled (from say, $500 to $2,000 over 100 years), the line would be twice as steep, reflecting a growth rate that was twice as high.

Interact

Follow figures click-by-click in the full interactive version at www.core-econ.org.

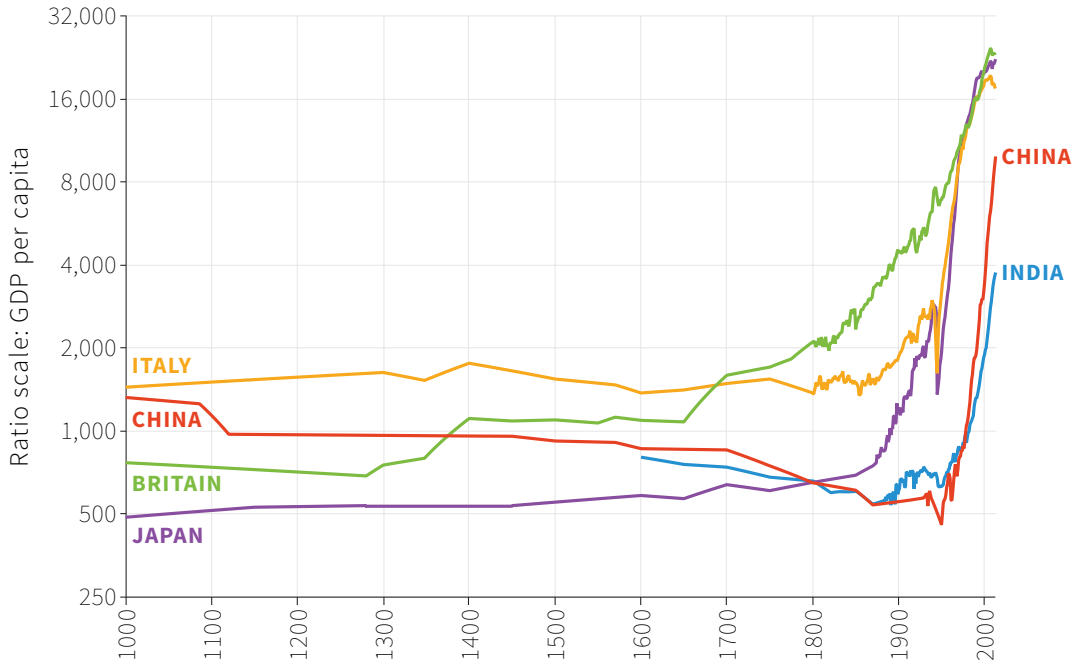


Figure 1.1b *History's hockey stick: Living standards in five countries (1000-2013) using the ratio scale.*

Source: *Bolt, Jutta, and Jan Juiten van Zanden. 2013. 'The First Update of the Maddison Project Re-Estimating Growth Before 1820.' Maddison-Project Working Paper WP-4. Broadberry, Stephen. 2013. 'Accounting for the Great Divergence.' London School of Economics and Political Science. November 1.*

History's hockey stick

There were cultural changes and scientific advances in many parts of the world over the entire period shown in the figure, but living standards began to rise in a sustained way only from the 18th century. The figure looks like a hockey stick, and our eyes are drawn to the kink. The hockey-stick kink is less abrupt in Britain, where growth began around 1650. In Japan the kink is more defined, occurring around 1870. The kink in China did not happen until around 1980, and in India even more recently. GDP per capita actually fell in India during British colonial rule. You can see that this is also true of China during the same period, when European nations dominated China's politics and economics. The ratio scale makes it possible to see that recent growth rates in Japan and China were higher than elsewhere.

If you have never have seen an ice-hockey stick (or ice hockey), this is why we call these figures *hockey stick curves*:

In some economies, substantial improvements in people's living standards did not occur until they gained independence from colonial rule or interference by European nations:

- When 300 years of British rule of India ended in 1947, according to Angus Deaton, an economist: "It is possible that the deprivation in childhood of Indians... was as severe as that of any large group in history". In the closing years of British rule, a child born in India could expect to live for 27 years. Fifty years on, life expectancy at birth in India had risen to 65 years.
- China had once been richer than Britain, but by the middle of the 20th century GDP per capita in China was one-fifteenth that of Britain.
- Neither Spanish rule of Latin America, nor its aftermath following the independence of most Latin American nations early in the 19th century saw anything resembling the hockey-stick upturn in living standards experienced by the countries in Figures 1.1a and 1.1b.

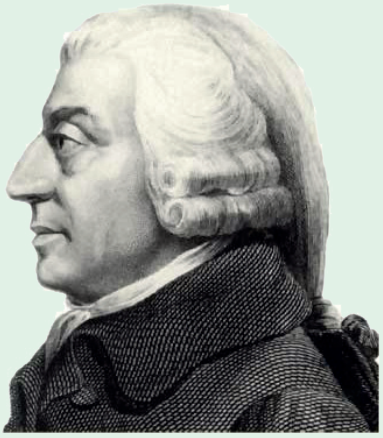
We learn two things from Figures 1.1a and 1.1b:

- For a very long time living standards did not grow in any sustained way.
- When sustained growth occurred it happened at different times in different countries, leading to vast differences between living standards around the world.

Understanding how this occurred has been among the most important questions that economists have asked, starting with the founder of the field, Adam Smith, who gave his most important book the title *An Inquiry into the Nature and Causes of the Wealth of Nations*.

GREAT ECONOMISTS

ADAM SMITH



Adam Smith (1723-1790), considered by many to be the father of economics, was raised by his widowed mother in Scotland. He studied philosophy at the University of Glasgow and later at Oxford, where he wrote: "the greater part of the… professors have… given up altogether even the pretence of teaching."

He travelled throughout Europe, visiting Toulouse, France where because he had "very little to do", he said, he had "begun to write a book in order to pass away the time." It became the most famous book in economics.

In An Inquiry into the Nature and Causes of the Wealth of Nations, published in 1776, Smith asked: how can society coordinate the independent activities of large numbers of economic actors—producers, transporters, sellers, consumers—often unknown to each other and widely scattered across the world? His radical claim was that coordination among all of these actors might spontaneously arise, without any person or institution consciously attempting to create or maintain it. This challenged previous notions of political and economic organisation, in which rulers imposed order on their subjects.

Even more radical was his idea that this could take place as a result of individuals pursuing their self interest: "It is not from the benevolence of the butcher, the brewer, or the baker that we expect our dinner, but from their regard to their own interest," he wrote, adding that each would be "led by an invisible hand to promote an end which was no part of his intention."

Since then this "invisible hand" has been a metaphor for how markets can coordinate the self-interested pursuits of people to produce a socially desirable outcome.

Smith did not think that people were guided entirely by self-interest, and he wrote a book about ethical behaviour called The Theory of Moral Sentiments, published in 1759.

He also understood that the market system had some failings, especially if sellers on markets banded together so as to avoid competing with each other. "People in the same trade seldom meet together," he wrote, "even for merriment and diversion, but the conversation ends in a conspiracy against the public; or in some contrivance to raise prices."

He specifically targeted monopolies that were protected by governments, such as the British East India Company that not only controlled trade between India and Britain, but also administered much of the British colony there.

He agreed with his contemporaries that government should protect the nation from external enemies and ensure justice through the police and the court system—he also advocated government investment in education, and in public works such as bridges, roads, and canals.

DISCUSS 1.1: THE ADVANTAGES OF CONVENTIONAL AND RATIO SCALES

Figure 1.1a used a conventional scale for the vertical axis, and Figure 1.1b used a ratio scale.

1. Choose any two of the countries shown in these figures and compare their growth from 1400 to the present, using the information in the figures.
2. Which figure is more helpful for this comparison and why?

1.2 MEASURING INCOME AND LIVING STANDARDS

The estimate of living standards, *GDP per capita*, that we used in Figures 1.1a and 1.1b is a measure of total income (and output) in a country (called *gross domestic product*, or *GDP*), which is then divided by the country's population.

GDP is a measure of the total output of the economy in a given period, such as a year: Diane Coyle, an economist, says it "adds up everything from nails to toothbrushes, tractors, shoes, haircuts, management consultancy, street cleaning, yoga teaching, plates, bandages, books, and the millions of other services and products in the economy".

Adding up these millions of services and products requires finding some measure of how much a yoga class is worth compared to a toothbrush. Economists must first decide what should be included, but also how to give a value to each of these things. In practice, the easiest way to do this is by using their prices.

Three important points to remember about measuring average living standards in a country:

- GDP is a measure of total income in a country; to get an average measure, GDP is divided by population, giving GDP per capita.
- GDP per capita is not the same as the *disposable income* of a typical person.
- A person's disposable income is a measure of his or her living standards, but it omits important aspects of wellbeing.

What do the second and third points mean? A person's living standard refers to how well off the person is. This is sometimes measured by an individual's disposable income. This is the amount of wages or salaries, profit, rent, interest and transfer payments from the government (such as unemployment or disability benefit) or from others (for example, gifts) received over a given period such as a year, minus any transfers the individual made to others including taxes paid to the government. Disposable income is thought to be a good measure of living standards because it is the maximum amount of food, housing, clothing and other goods and services that the person can buy without having to borrow—that is, without going into debt or selling possessions. But, if your disposable income was used to represent your living standard, you might question this for two reasons:

- Is our disposable income a good measure of our wellbeing?
- When we're part of a group of people (a nation for example, or an ethnic group) is the average disposable income a good measure of how well off the group is?

Disposable income and wellbeing

Income is a major influence on wellbeing because it allows us to buy the goods and services that we need or enjoy. But it is insufficient, because many aspects of our wellbeing are not related to what we can buy. For example, disposable income leaves out:

- The quality of our social and physical environment such as friendships and clean air.
- Goods and services that we do not buy, such as healthcare and education if they are provided by a government.
- Goods and services that are produced within the household, such as meals or childcare (predominantly provided by women).

Average disposable income and average wellbeing

Consider a group of people in which each person initially has a disposable income of $5,000 a month, and imagine that, with no change in prices, income has risen for every individual in the group. Then we would say that average or typical wellbeing had risen.

But now think about a different comparison. In a second group, the monthly disposable income of half the people is $10,000. The other half has just $500 to spend every month. The average income in the second group ($5,250) is higher than in the first (which was $5,000 before incomes rose). But would we say that the second group's wellbeing is greater than that of the first group, where everyone has $5,000 a month? The additional income in the second group is unlikely to matter much to the rich people, but the poor half would think their poverty was a serious deprivation.

Absolute income matters for wellbeing, but we also know from research that people care about their relative position in the income distribution. They report lower wellbeing if they find they earn less than others in their group.

Since income distribution affects wellbeing, and because the same average income may result from very different distributions of income between rich and poor within a group, average income may fail to reflect how well off a group of people is by comparison to some other group.

Valuing government goods and services

GDP includes the goods and services produced by the government, such as schooling, national defence, and law enforcement. They contribute to wellbeing but are not included in disposable income. In this respect, GDP per capita is a better measure of living standards than disposable income.

But government services are difficult to value, even more difficult to value than services such as haircuts and yoga lessons. For goods and services that people buy we take their price as a rough measure of their value (if you valued the haircut less than its price, you would have just let your hair grow). But the goods and services produced by government are typically not sold, and the only measure of their value to us is how much it cost to produce them.

The gaps between what we mean by wellbeing, and what GDP per capita measures, should make us cautious about the literal use of GDP per capita to measure how well off people are. But when the changes over time or differences among countries in this indicator are as great as those in Figures 1.1a and 1.1b (and in Figures 1.9 and 1.10 later in this unit), GDP per capita is undoubtedly telling us *something* about the differences in the availability of goods and services.

We look in more detail at how GDP is calculated so that we can compare it through time, and make comparisons between countries, in this unit's Einstein section (many of the units have Einstein sections: they will show you how to calculate many

of the statistics that we use). Using these methods, we can use GDP per capita to unambiguously communicate such ideas as "people today in Japan are on average a lot richer than they were 200 years ago, and a lot richer than the people of India today."

Looking at the two parts of Figure 1.1, the obvious next question is: what changed so dramatically in the past 300 years?

DISCUSS 1.2: WHAT SHOULD WE MEASURE?

While campaigning for the US presidency on 18 March 1968, Senator Robert Kennedy gave a famous speech questioning "the mere accumulation of material things" in American society, and why, among other things, air pollution, cigarette advertising and jails were counted when the US measured its living standards, but health, education or devotion to your country were not. He argued that: "It measures everything, in short, except that which makes life worthwhile."

Read his speech in full, or listen to a sound recording of it.

3. In the full text, which goods does he list as being included in a measure of GDP?
4. Do you think these should be included in such a measure, and why?
5. Which goods does he list in the full text as missing from the measure?
6. Do you think they should be included, and why?

1.3 THE PERMANENT TECHNOLOGICAL REVOLUTION

Remarkable scientific and technological advances occurred more or less at the same time as the upward kink in the hockey stick in Britain in the middle of the 18th century. Important new technologies were introduced in textiles, energy and transportation. Its cumulative character led to it being called the *Industrial Revolution*.

As late as 1800, traditional craft-based techniques, using skills that had been handed down from one generation to the next, were still used in most production processes. The new era brought new ideas, new discoveries, new methods and new machines, making old ideas and old tools obsolete. These new ways were, in turn, made obsolete by even newer ones.

Although in everyday usage, *technology* refers to machinery, equipment and devices developed using scientific knowledge, in economics, technology is a process that takes a set of materials and other inputs—including the work of people and machines—and creates an output. For example, a technology for making a cake can be described by the recipe that specifies the combination of inputs (ingredients such as flour, and labour activities such as stirring) needed to create the output (the cake). Another technology for making cakes uses large-scale machinery, ingredients and labour (machine operators).

Until the Industrial Revolution, the economy's technology, like the skills needed to follow its recipes, was updated only slowly and passed from generation to generation. As *technological progress* revolutionised production, the time required to make a pair of shoes fell by half in only a few decades; the same was true of spinning and weaving, and of making cakes in a factory. This marked the beginning of a permanent technological revolution because the amount of time required for producing most products fell generation after generation.

Technological change in lighting

To get some idea of the unprecedented pace of change, consider the way we produce light. For most of human history technological progress in lighting was slow. Our distant ancestors typically had nothing brighter than a campfire at night. The recipe for producing light (had it existed) would have said: gather lots of firewood, borrow a lighting stick from some other place where a fire is maintained, and start and maintain a fire.

The first great technological breakthrough in lighting came 40,000 years ago, with the use of lamps that burned animal or vegetable oils. We measure technological progress in lighting by how many units of brightness called *lumens* could be generated by an hour of work. One lumen is approximately the amount of brightness in a square metre of moonlight. One lumen-hour (lm-hr) is this amount of brightness lasting an hour. For example, creating light by a campfire took about 1 hour of labour to produce 17 lm-hr, but animal fat lamps produced 20 lm-hr for the same amount of work. In Babylonian times (1750 BC) the invention of an improved lamp using sesame oil meant that an hour of labour produced 24 lm-hr. Technological progress was slow: this modest improvement took 7,000 years.

Three millennia later, in the early 1800s, the most efficient forms of lighting (using tallow candles) provided about nine times as much light for an hour of labour as had the animal fat lamps of the past. Since then lighting has become more and more efficient with the development of town gas lamps, kerosene lamps, filament bulbs,

fluorescent bulbs and other forms of lighting. Compact fluorescent bulbs introduced in 1992 are about 45,000 times more efficient, in terms of labour time expended, than lights were 200 years ago. Today the productivity of labour in producing light is half a million times greater than it was among our ancestors around their campfire.

Figure 1.2, below, charts this remarkable hockey-stick growth in efficiency in lighting using the ratio scale we introduced in Figure 1.1b.

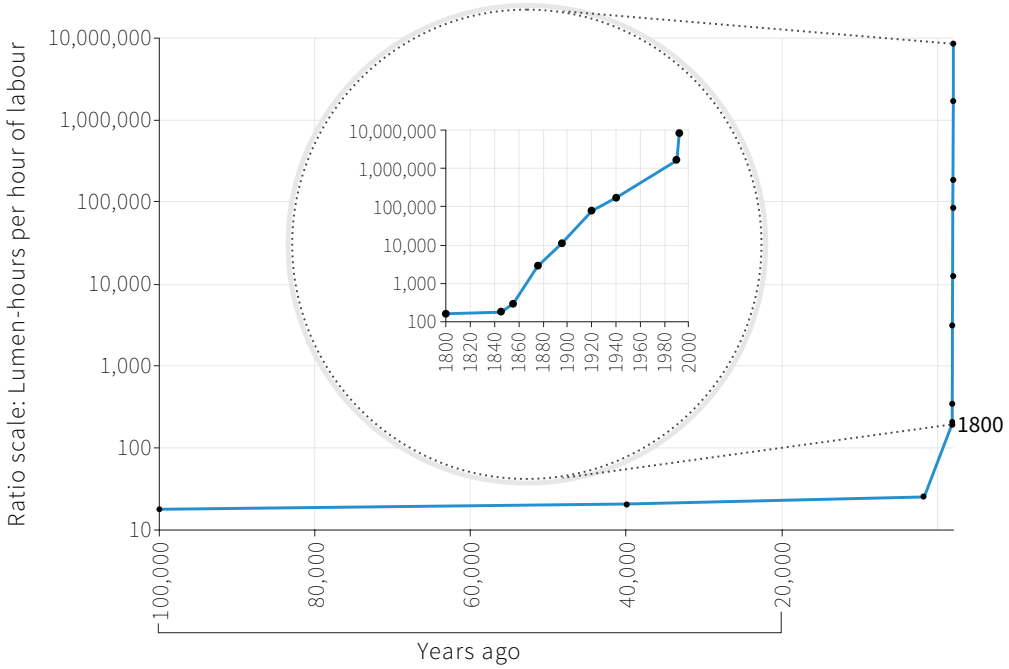


Figure 1.2 *The productivity of labour in producing light: Lumen-hours per hour of labour (100,000 years ago to the present).*

Source: *Nordhaus, William. 1998. 'Do Real Output and Real Wage Measures Capture Reality? The History of Lighting Suggests Not.' Cowles Foundation For Research in Economics Paper 957.*

The process of innovation did not end with the Industrial Revolution as the case of labour productivity in lighting shows. It has continued with the application of new technologies in many industries, such as the steam engine, electricity, transportation (canals, railroads, automobiles), and most recently, the revolution in information processing and communication. These broadly applicable technological innovations give a particularly strong impetus to growth in living standards because they change the way large parts of the economy work.

By reducing the amount of work time it takes to produce the things we need, technological changes allowed significant increases in living standards. David Landes, an economic historian, wrote that the Industrial Revolution was "an interrelated succession of technological changes" that transformed the societies in which these changes took place. This process continues today: Hans Rosling, a statistician, claims, in this video of a TED lecture, that we should say "thank you

industrialisation" for creating the washing machine, a labour-saving device that had a far-reaching effect on the wellbeing of millions of women, including his own mother.

1.4 A CONNECTED WORLD

In July 2012 the Korean hit Gangnam Style was released. By the end of 2012 it had been the best-selling song in 33 countries, including Australia, Russia, Canada, France, Spain and the UK. With 2 billion views by the middle of 2014, Gangnam Style also became the most watched video on YouTube. The permanent technological revolution has produced a connected world.

Gagnam Style

Everyone is part of it. The materials making up this introduction to economics were written by teams of economists, designers, programmers and editors, working together—often simultaneously—at computers in the UK, India, the US, Russia, Colombia, South Africa, Chile, Turkey, France and many other countries. If you are online, some of the transmission of information occurs at close to the speed of light. While most of the commodities traded around the globe still move at the pace of an ocean freighter, about 21 miles (33km) per hour, international financial transactions are implemented in less time than it took you to read this sentence.

The speed at which information travels provides more evidence of the novelty of the permanent technological revolution. By comparing the known date of a historical event with the date at which the event was first noted in other locations (in diaries, journals or newspapers) we can determine the speed at which news travelled. When Abraham Lincoln was elected US President in 1860, for example, the word was spread by telegraph from Washington to Fort Kearny, which was at the western end of the telegraph line. From there the news was carried by a relay of riders on horseback called the Pony Express, covering 1,260 miles (2,030km) to Fort Churchill in Nevada, from where it was transmitted to California by telegraph. The process took seven days and 17 hours. Over the Pony Express segment of the route, the news travelled at 7 miles (11km) per hour. A half-ounce (14 gram) letter carried over this route cost $5, or the equivalent of five days' wages.

From similar calculations we know that news travelled between ancient Rome and Egypt at about 1 mile (1.6km) per hour, and 1,500 years later between Venice and other cities around the Mediterranean it was, if anything, slightly slower. But, a few

centuries later, as Figure 1.3 shows, the pace began to quicken. It took "only" 46 days for the news of a mutiny of Indian troops against British rule in 1857 to reach London, and readers of the *Times* of London knew of Lincoln's assassination only 13 days after the event. One year after Lincoln's death a transatlantic cable cut the time for news to travel between New York and London to a matter of minutes.

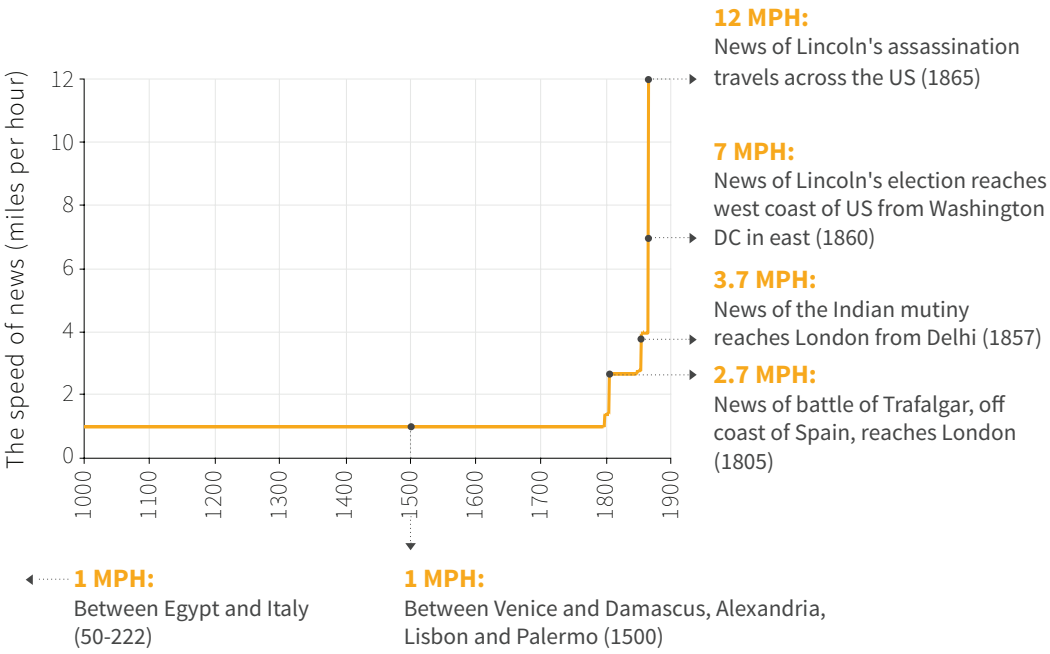


Figure 1.3 *The speed at which information travelled (1000 to 1865).*

Source: *Tables 15.2 and 15.3 from Clark, Gregory. 2007. A Farewell to Alms: A Brief Economic History of the World. Princeton, NJ: Princeton University Press.*

1.5 THE GROWTH OF POPULATION AND CITIES

Alongside technological progress and a rising standard of living, population has grown rapidly. For most of the last 12,000 years the population of the world grew slowly, if at all, with increases in good years followed by declines in response to climatic adversity and other disasters.

Figure 1.4 shows the evolution of world population from the year 1000 onwards. In a few countries, population started to grow rapidly 200 years ago, but the world's population took off in the 20th century with the development and spread of improved sewerage, clean water, and other public health measures. While the number of people in the world continues to grow, as shown in Figure 1.4, the pace of growth is slowing from its peak in the 1970s (see Figure 1.5). The *demographic*

transition refers to the slowdown in population growth as the fall in death rates is balanced by a fall in birth rates associated with the desire for fewer children, combined with public policies discouraging larger families, as in China.

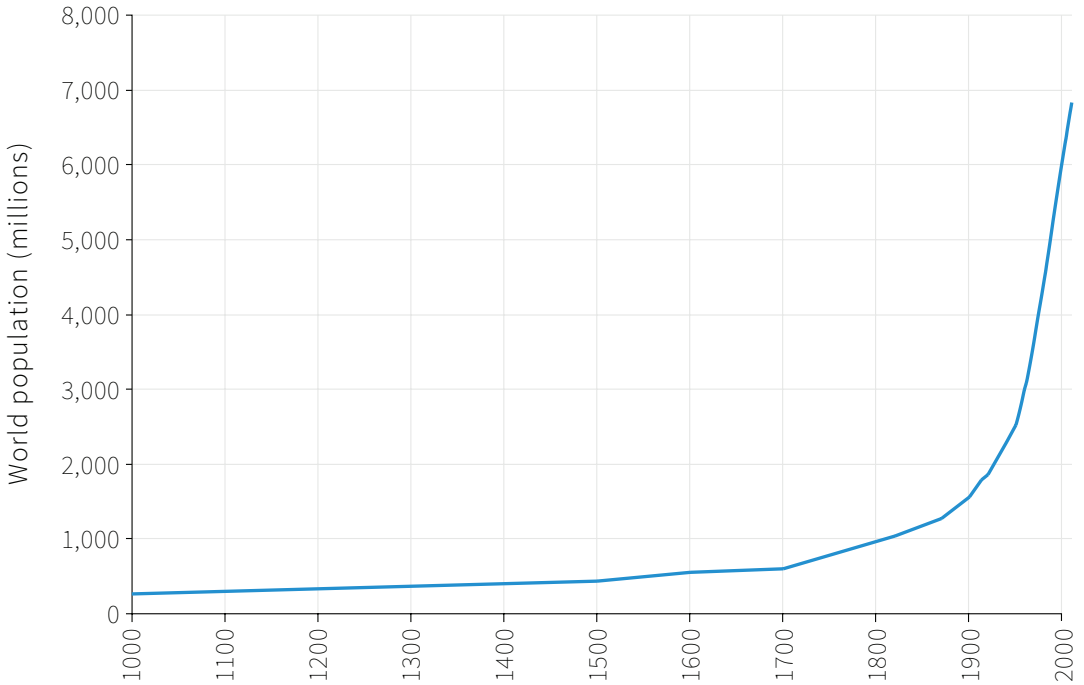


Figure 1.4 *World population (1000-2010).*

Source: *Maddison, Angus. 2015. 'Statistics on World Population, GDP and Per Capita GDP, 1-2008 AD.' Accessed June 2015, and US Census Bureau. 2015. 'International Programs, International Data Base.' Accessed June 2015.*

With the increased productivity of labour in agriculture, fewer farmers were required to feed the nonfarming population. Higher labour productivity means that on a given piece of land, more output could be produced by each farmer. People left farming to pursue other occupations, resulting in another change: the growth of cities. Three hundred years ago, the vast majority of people lived in the countryside interacting with just a handful of people, mostly family and neighbours. In the last few centuries, however, people have been drawn—or, in some cases, pushed—into cities.

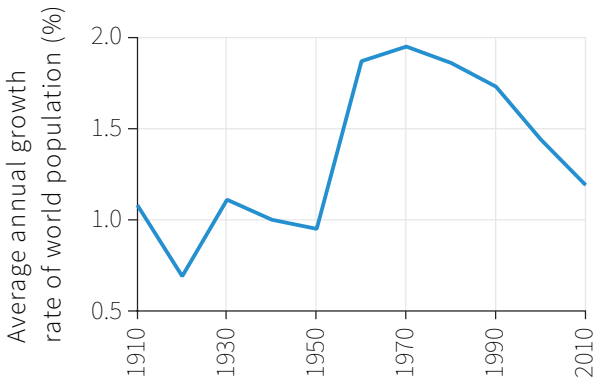


Figure 1.5 *How the world's population growth in the 20th century rose and fell.*

Source: *Angus Maddison historical statistics. US Census: World population growth rate.*

DISCUSS 1.3: WORLD POPULATION DATA

Click on the source of Figure 1.4 and then on the link called "statistics on world population". This will automatically download an Excel file to your computer with, among other data, the data used to plot Figure 1.4 (this is the second worksheet in the Excel file).

Scroll down to the rows showing China and India and add up these numbers to show a total population for China and India for each year.

1. Plot this total population in a graph similar to Figure 1.4. Now insert the total population for the 30 Western European countries into the same graph. What can you say about population growth in these two groups of countries over time?
2. Finally, use this data to plot the ratio scale version of this graph (refer to the description of a ratio scale in Section 1.1). Compare the population growth rate of these two groups of countries using your new graph. Can you explain the differences in the growth rates?
3. What are the implications of the differences in (2)?

City living is a drastic change, as everyday life is populated by dozens or even hundreds of strangers. This of course changes how we interact with others—many of whom we will never see again—in some cases challenging people's personal security and requiring new ways of maintaining social order. Police forces are a relatively new feature of human society, dating from the emergence of large urban areas.

In 1850 there were only three cities with populations exceeding 1 million people—London, Paris, and Beijing—but, as Figure 1.6 demonstrates, by 2013 there were more than 500 cities of this size.

Tokyo, the world's biggest urban area, is home to 34 million people. That's four times as many people living in one city today as archaeologists think existed in the entire world 11,000 years ago, at the time humans first took up farming. In 1900, nine of the 10 largest cities in the world were in Europe or North America—Tokyo was the exception. Today nine of the 10 are in Asia or Latin America, with New York the odd one out.

Tokyo: Birds-eye view

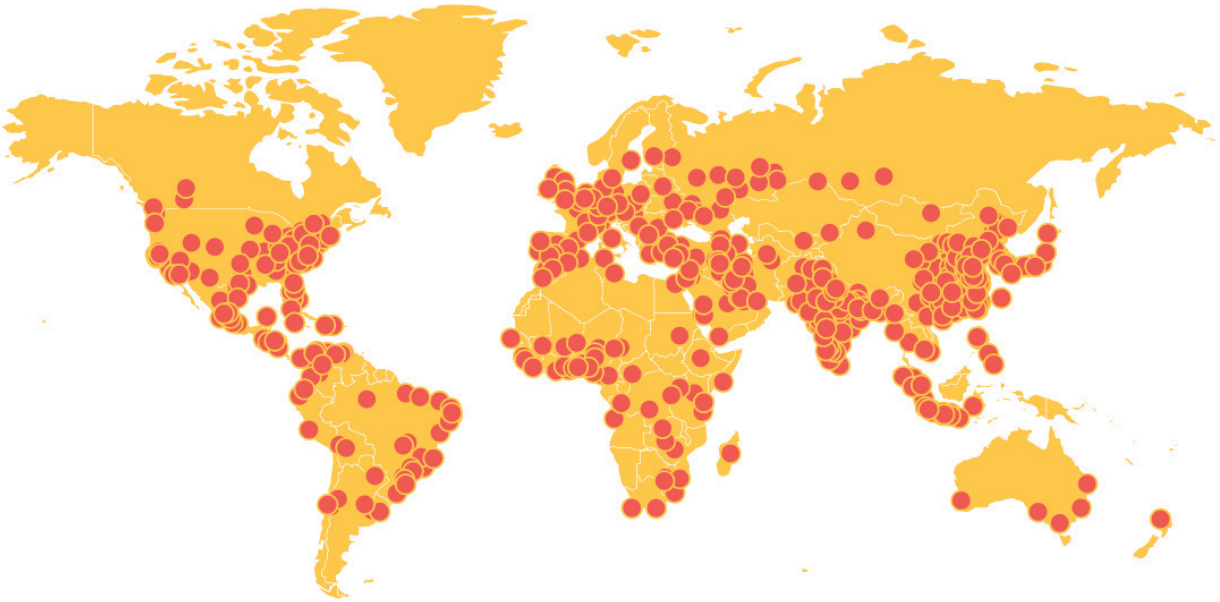


Figure 1.6 *Cities with more than 1 million inhabitants (2013).*

Source: *'Major Agglomerations of the World - Population Statistics and Maps.' 2015. Accessed June 2015. Data is for agglomerations (a central city and neighbouring towns (suburbs) forming a connected region of dense, predominately urban population with more than 1 million inhabitants.*

1.6 IMPACTS ON THE ENVIRONMENT

As production has soared (Figures 1.1a and 1.1b, and also Figure 1.2), so too have both the use of our natural resources and degradation of our natural environment. Elements of the ecological system such as air, water, soil, and weather have been altered by humans more radically than at any time in human history.

One example is climate change. Figures 1.7a and 1.7b present evidence that our use of fossil fuels—coal, oil, and gasoline—have profoundly affected the natural environment. After having remained relatively unchanged for many centuries, increasing emissions of carbon dioxide (CO_2) into the air during the 20th century have resulted in measurably larger amounts of CO_2 in the earth's atmosphere (Figure 1.7a) and brought about perceptible increases in the northern hemisphere's average temperatures (Figure 1.7b). Figure 1.7a also shows that CO_2 emissions from fossil fuel consumption have risen dramatically over the past 250 years.

DISCUSS 1.4: THE ENVIRONMENTAL KUZNETS CURVE

Many researchers think that there is a hump-shaped relationship between a country's income and environmental degradation. This relationship is often referred to as the *Environmental Kuznets Curve (EKC)*.

1. Read this description of the EKC and, in your own words, explain why such a relationship might be observed.
2. How might this relationship change when we define income as GDP versus GDP per capita?

Figure 1.7b shows that the average temperature of the earth fluctuates from decade to decade. Many factors cause these fluctuations, including volcanic events such as the Mount Tambora eruption in Indonesia, in 1815. Mount Tambora spewed so much ash that the Earth's temperature was reduced, and 1816 became known as the "year without a summer".

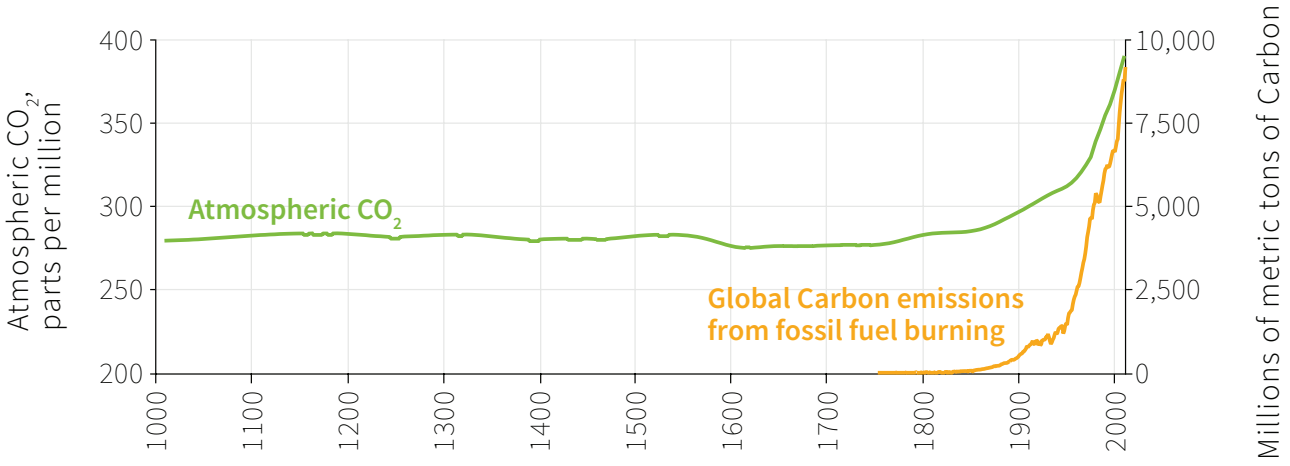


Figure 1.7a *Carbon dioxide in the atmosphere (1010-2010) and global carbon emissions from burning fossil fuel (1750-2010).*

Source: *Years 1010-1975: Etheridge, D. E., L. P. Steele, R. J. Francey, and R. L. Langenfelds. 2012. 'Historical Record from the Law Dome DE08, DE08-2, and DSS Ice Cores.' Division of Atmospheric Research, CSIRO, Aspendale, Victoria, Australia. Years 1976-2010: Data from Mauna Loa observatory. Boden, T. A., G. Marland, and R. J. Andres. 2010. 'Global, Regional and National Fossil-Fuel CO2 Emissions.' Carbon Dioxide Information Analysis Center (CDIAC) Datasets.*

In the last century, average temperatures have risen in response to increasingly high levels of greenhouse gas concentrations. These have resulted from the CO2 emissions associated with the burning of fossil fuels. The likely consequences of global warming are far-reaching: melting of the polar ice caps, rising sea levels that may put large coastal areas under water, and potential changes in climate and rain patterns that may destroy the world's food-growing areas.

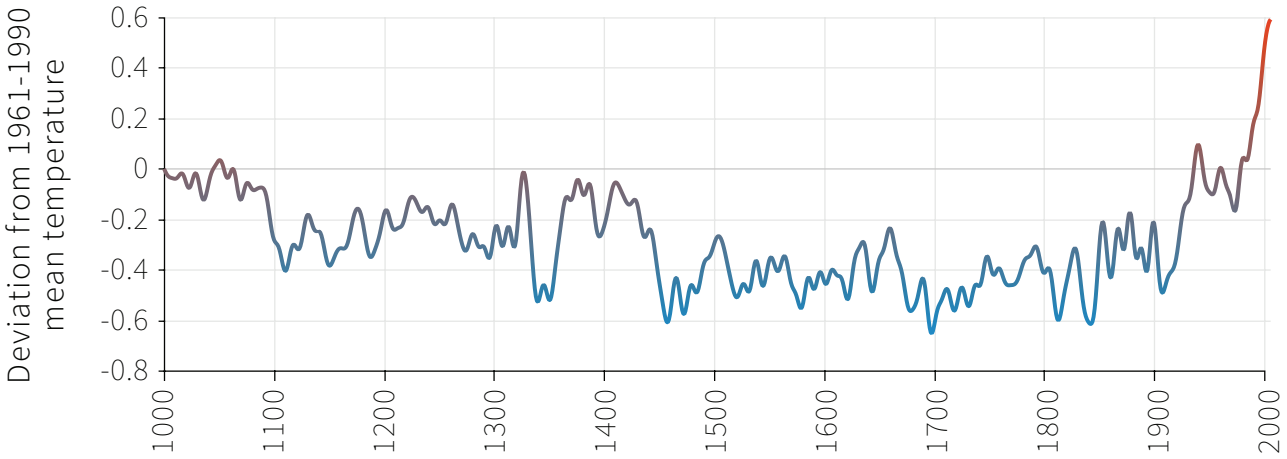


Figure 1.7b *Northern hemisphere temperature over the long run (1000-2006).*

Source: *Mann, M. E., Z. Zhang, M. K. Hughes, R. S. Bradley, S. K. Miller, S. Rutherford, and F. Ni. 2008. 'Proxy-Based Reconstructions of Hemispheric and Global Surface Temperature Variations over the Past Two Millennia.' Proceedings of the National Academy of Sciences 105 (36): 13252–57.*

Climate change is a global development. But many environmental impacts are local, as residents of cities suffer respiratory and other illnesses as a result of high levels of harmful emissions from power plants, vehicles, and other sources. Rural communities, too, are impacted by deforestation and the depletion of the supply of clean water and fishing stocks.

These examples of the way that people affect and are affected by both local and global ecologies motivate the way that we use the word "economy". When we named our ebook The Economy we were thinking about the way that people interact with each other, and also with nature, in producing their livelihood.

Figure 1.8 shows one way of thinking about the economy: the economy is part of a larger social system, which is itself part of the biosphere, which is the collection of all forms of life on earth.

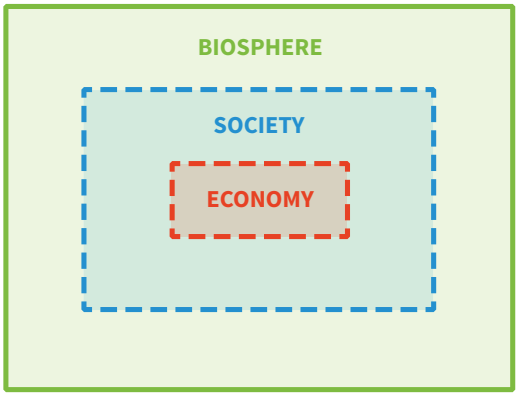


Figure 1.8 *The economy is part of society, which is part of the biosphere.*

From global climate change to local resource exhaustion, these effects are results of both the expansion of the economy (illustrated by the growth in total output) and the way the economy is organised (what kinds of things are valued and conserved, for example).

There is no doubt that the permanent technological revolution—which brought about dependence on fossil fuels—is part of today's environmental problem. But it is also part of the solution.

Look back at Figure 1.2, which showed the productivity of labour in producing light. The vast increases shown over the course of history and especially since the mid 19th century occurred in large part because the amount of light produced per unit of heat (for example from a campfire, candle, or light bulb) increased dramatically.

In lighting, the permanent technological revolution brought us more light for less heat, which conserved natural resources—from firewood to fossil fuels—used in generating the heat. Advances in technology today may allow greater reliance on wind, solar and other renewable sources of energy.

CLIMATE CHANGE

The human causes, and the reality, of climate change are no longer widely disputed in the scientific community.

The Intergovernmental Panel on Climate Change is the authoritative source for research and data. The likely consequences of global warming are far-reaching: melting of the polar ice caps, rising sea levels that may put large coastal areas under water, and potential changes in climate and rain patterns that may destroy the world's food-growing areas. The long-term physical and economic consequences of these changes, and the appropriate policies that governments could adopt as a result, are discussed in detail in Unit 18.

1.7 CAPITALISM DEFINED

Looking back over the data in Figures 1.1 to 1.7 we see an upward turn, like the kink in our hockey stick, repeated for:

- Gross domestic product per capita
- Productivity of labour (light per hour of work)
- Connectivity of the various parts of the world (the speed at which news travels)
- World population
- Impact of the economy on the global environment (Carbon emissions, atmospheric CO_2 and climate change)

How can we explain the change from a world in which living conditions fluctuated if there was an epidemic or a war, to a situation in which most of the time each generation is noticeably, and predictably, better off than the previous one?

The answer that makes most sense both factually and logically is what we call the capitalist revolution, which introduced a new *economic system* called *capitalism* characterised by a new combination of *institutions*. An economic system is a way of organising the production and distribution of goods and services in an entire economy. And by institutions, we mean the different sets of laws and social customs regulating production and distribution in different ways in families, private businesses, and government bodies.

CAPITALISM

An economic system in which three key institutions play an important role:

- Private property
- Markets
- Firms

In some economies in the past the key economic institutions were private property, markets and families, because production usually took place in families rather than firms. Think about a farm owned by a family: who does the work? Who consumes the produce? This has historically been determined by the older generation of the family (in most societies, the older men), and by social custom.

In other societies the government has been the institution governing production, distribution and the process of change. In this case, most production has taken place in government-owned establishments, and the government has decided how the goods that were produced were used, including who gets what. This is called a *centrally planned* economic system. It existed, for example, in the Soviet Union, East Germany and many other eastern European countries prior to the end of Communist Party rule in the early 1990s.

Though governments and families are essential parts of the workings of every economy, most economies today are capitalist. Since most of us live in capitalist economies, it is easy to overlook the importance of institutions that are fundamental for capitalism to work well: they are so familiar, we hardly ever notice them. Before seeing how private property, markets and firms combine in the capitalist economic system, we need to define them.

1.8 PRIVATE PROPERTY, MARKETS AND FIRMS

PRIVATE PROPERTY

Private property means that you can:

- Enjoy your possessions in a way that you choose
- Exclude others from their use if you wish
- Dispose of them by gift or sale to someone else...
- ... who becomes their owner

Over the course of human history the extent of private property has varied. In some societies, such as the hunters and gatherers who are our distant ancestors, almost nothing except personal ornaments and clothing was owned by individuals. In others, crops and animals were private property, but land was not. The right to use the land was granted to families by consensus among members of a group, or by a chief, without allowing the family to sell the plot.

In other economic systems other human beings—slaves—were private property.

In a capitalist economy, an important form of private property is made up of the equipment, buildings, raw materials, patents and other intellectual property, and other inputs used in producing goods and services. These are called *capital goods*.

In a capitalist economy, private property does not include some essentials such as the air we breathe and most of the knowledge we use (such as our skills, our knowledge of how to produce things, and our capacities to solve problems that arise in production). Private property may be owned by an individual, a family, a business, or some entity other than the government.

Think of all the ways that goods and services may be transferred from one person to another: as a gift, by theft, by a government order. *Markets* differ from these, and other ways that goods or services may be transferred from one person to another, in two respects:

- *They are reciprocated*: First, unlike gifts and theft, in a market one person's transfer of a good or service to another is directly

MARKETS

Markets are:

- A way of connecting people who mutually benefit
- By exchanging goods and services
- Through a process of buying and selling

reciprocated by a transfer in the other direction (either of another good or service as takes place in barter exchange, or money, or a promise for a later transfer when one buys on credit).

- *They are voluntary*: Both transfers—by the buyer and the seller—are voluntary because the things being exchanged are private property. So the exchange must be beneficial in the opinion of both parties. In this, markets differ from theft, and also from the transfers of goods and services in a centrally planned economy.

DISCUSS 1.5: THE POOREST MAN'S COTTAGE

"The poorest man may in his cottage bid defiance to all the forces of the Crown. It may be frail, its roof may shake; the wind may blow through it; the storms may enter, the rain may enter—but the King of England cannot enter; all his forces dare not cross the threshold of the ruined tenement."
William Pitt, 1st Earl of Chatham, *speech in the British Parliament* (1763)

1. What does this tell us about the meaning of private property?
2. Does it apply to people's homes in your country?

DISCUSS 1.6: MARKETS AND SOCIAL NETWORKS

Think about a social networking site that you use, for example Facebook. Now look at our definition of a market.

What are the similarities and differences between the social networking site and a market?

But private property and markets alone do not define capitalism. In many places they had been important institutions long before capitalism. The most recent of the three components making up the capitalist economy is the *firm*.

The kinds of firms that make up a capitalist economy include restaurants, banks, large farms that pay others to work there, industrial establishments, supermarkets, internet service providers, and many more. Other productive organisations that are not firms and which play a lesser role in a capitalist economy include family

businesses, in which most or all of the people working are family members, non-profit organisations, employee-owned cooperatives, and government-owned entities (such as railways and power or water companies). These are not firms, either because they do not make a profit, or because the owners are not private individuals who own the assets of the firm and employ others to work there. Note: a firm pays wages or salaries to employees; but if it takes on unpaid student interns, it is still a firm.

Firms existed, playing a minor role, in many economies long before they became the predominant organisations for the production of goods and services, as they are in a capitalist economy. This created a boom in another kind of market that had played a limited role in earlier economic systems: the *labour market*. The owners of the firms as employers (or their managers) offer jobs at wages or salaries that are high enough to attract people who are looking for work.

FIRM

A *firm* is a way of organising production with the following characteristics:

- One or more individuals own a set of capital goods that are used in production
- They pay wages and salaries to employees
- They direct the employees (through the managers they also employ) in the production of goods and services
- The goods and services are the property of the owners
- Who sell them on markets with the intention of making a profit

In economic language, the employers are the *demand side* of the labour market (they "demand" employees), while the workers are the *supply side*, offering to work under the direction of the owners and managers who hire them.

A striking characteristic of firms, that distinguishes them from families and governments, is how quickly they can be born, expand, contract and die. A successful firm can grow from just a few employees to a global company with hundreds of thousands of customers, employing thousands of people, in a few years. Firms can do this because they are able to hire additional employees on the labour market, and attract funds to finance the purchase of the capital goods they need to expand production.

Firms can die in a few years too. This is because a firm that does not make profits will not have enough money (and will not be able to borrow money) to continue employing and producing. The firm shrinks, and some of the people who work there lose their jobs.

Contrast this with a successful family farm. The family will be better off than its neighbours; but unless it turns the family farm into a firm, and employs other people to work on it, expansion will be limited. If, instead, the family is not very good at farming, then it will simply be less well off than its neighbours. The family head cannot make his children redundant. As long as the family can feed itself there is no equivalent mechanism to a firm's failure that will automatically put it out of business.

Government bodies tend to be more limited in their capacity to expand if successful, and are usually protected from failure if they perform poorly.

Markets and private property are essential parts of how firms function for two reasons:

- *Inputs and outputs are private property*: The firm's buildings, equipment, patents, and other inputs into production, as well as the resulting outputs, belong to the owners.
- *Firms use markets to sell outputs*: The owners' profits depend on markets in which customers may willingly purchase the products at a price that will more than cover their costs.

One way to remember the distinctiveness of the *capitalist* economic system is that unlike other economic systems, one of its hallmarks is the private ownership of *capital goods* that are organised for use in firms. Other economic systems are distinctive because of the importance of privately owned land, the presence of slaves, because the government owns capital goods, or because of the limited role of firms. Capitalist economies differ, too, from earlier economies in the magnitude of the capital goods used in production. Massive power looms have replaced spinning wheels; a tractor now pulls a plough to do a job once done by a farmer using a hoe.

1.9 CAPITALISM AS AN ECONOMIC SYSTEM

Figure 1.9 shows that the three parts of the definition of a capitalist economic system are nested concepts. Private property is an essential condition for the operation of markets, and the firm, in turn, presupposes markets and private property. The left-hand circle describes an economy of isolated families who own their capital goods and the goods they produce, but have little or no exchange with others.

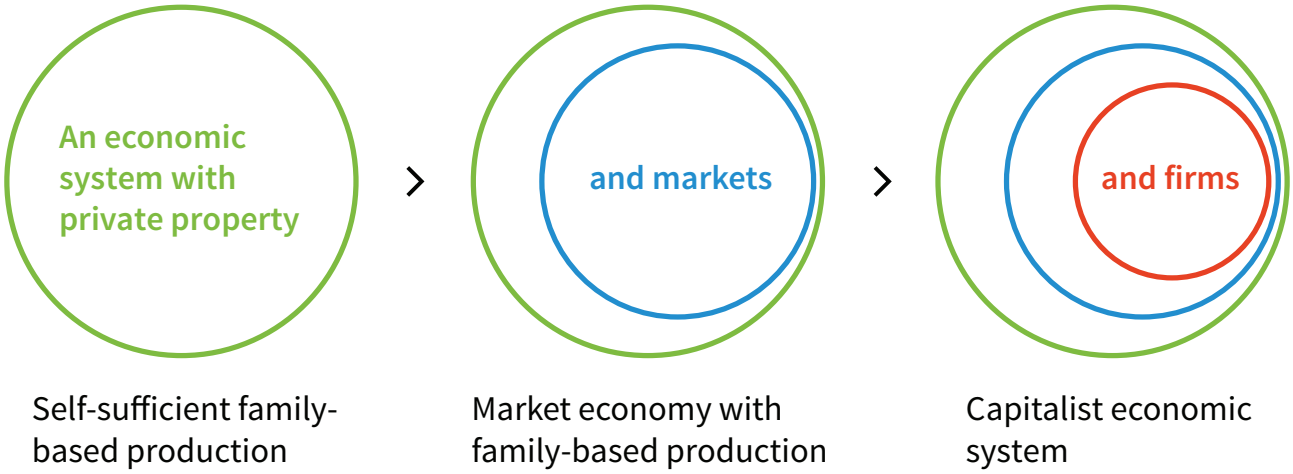


Figure 1.9 *Capitalism: Private property, markets and firms.*

Historically, economies like the left-hand circle have existed, but have been much less important than a system in which markets and private property are combined (the middle circle). In the middle circle most production takes place either by individuals (shoemakers or blacksmiths, for example) or in families (in our example, this was on a farm). Prior to 1600 a great many of the economies of the world were like this.

Capitalism is an economic system that combines decentralisation with centralisation:

- *Capitalism decentralises*: It limits the powers of governments and of other individuals in the process of owning, buying and selling.
- *Capitalism centralises*: It concentrates power in the hands of owners and managers of firms who are then able to secure the cooperation of large numbers of employees in the production process.

An easy way to remember this contrast is that when the owner of a firm interacts with an employee, he or she is "the boss". When the same owner interacts with a potential customer he or she is simply another person trying to make a sale, in competition with other firms. It is this unusual combination of competition among firms, and concentration of power and cooperation within them, that accounts for capitalism's success as an economic system.

How the institutions of capitalism—private property, markets, and firms—combine with each other and with families, governments, and other institutions differs greatly across countries. Just as ice and steam are both water, China and the US are both capitalist economies. But they differ in the extent to which the government influences economic affairs, and in many other ways. As this demonstrates, definitions in the social sciences often cannot be as precise as they are in the natural sciences.

Learning a new language

We hope you will not only learn about the economy in this course but also learn to *do* economics, and this means learning to speak a new language. Using the terms of economics helps us to communicate complicated ideas with others who have learned the language. This is why we stress definitions.

Being able to explain how economists use words is also crucial to communicating with other people about economics. For this reason, and because by now you have seen a number of definitions, think about what a definition does for us.

Water, for example, is defined chemically as a compound of two hydrogen atoms bonded with one oxygen atom, which takes the liquid form but also a solid form (ice) and a gaseous form (steam), not to mention other forms (snow or fog). Some people might say that "ice is not *really* water", and object that the definition is not the "true meaning" of the word.

But debates about "true" meaning (especially referring to complex ideas like capitalism, or democracy) misunderstand why definitions are valuable. Think of the definition of water, or of capitalism, not as capturing some true meaning—but rather as a device that is valuable because it makes it easier to communicate.

The word "capitalism", like "water", refers not to a single thing, but to a class of things sharing common characteristics. And, like the definition of water (which requires that we know how to use the words oxygen and hydrogen precisely), we needed to define the three institutions making up the capitalist economic system before we could define capitalism itself.

But unlike water, we cannot identify a capitalist economic system using easy-to-see physical characteristics.

Britain was definitely capitalist in 1800 and definitely not capitalist in 1500, but it would be pointless to try to find a precise date at which a switch occurred. For much of the period of transition we would say that the economy was a mixed economic system with both capitalist and non-capitalist elements.

China was a centrally planned economy from 1953 until economic reforms began in 1978. Afterwards it adopted new institutions so that markets, private property and firms became important. Today it is a capitalist economy. But in which year, exactly, did the definition come to be justified?

Major distinctions are important—the difference between a centrally planned and a capitalist economy, for example—but we can admit that the boundary between one and the other is rarely precise in real life, and so the way we describe a system will always be subjective. Even today, although capitalism is dominant in China, there is still a centrally organised Five Year Plan.

DISCUSS 1.7: FIRM OR NOT?

Using our definition, explain whether each of the following entities is a firm by stating whether it *satisfies the characteristics that define a firm*. Research the entity online if you are stuck.

1. John Lewis (UK)
2. A family farm in Vietnam
3. Your current family doctor's office or practice
4. Walmart (US)
5. An 18th century pirate ship (see our description of *The Rover* in Unit 5)
6. Google (US)
7. Manchester United plc (UK)
8. Wikipedia

1.10 CAPITALISM, CAUSATION AND HISTORY'S HOCKEY STICK

There are both historical and logical reasons for thinking that the emergence of capitalism as an economic system is one of the causes of the upward kink in the hockey sticks we have seen.

But we should be sceptical when anyone claims that something complex (capitalism) "causes" increased living standards (or technological improvement, population growth, a networked world, or environmental challenges).

In science, we support the statement that *X* causes *Y* by:

- *Understanding the relationship* between cause (*X*) and effect (*Y*)
- *Performing experiments* to gather evidence by measuring *X* and *Y*

In physics, we have a good understanding of how heat changes the state of water (transforming some of it to steam, for example), and we can easily do an experiment to see what happens when we raise its temperature to 100C (you repeat this experiment whenever you boil water). Therefore we can make a convincing causal statement about what will happen when we raise the temperature of water.

Equivalent causal statements are essential in economics. We would often like to devise ways of changing something so that the economy works better, and this means making a causal statement that policy X is likely to cause change Y. For example, an economist might claim that: "If the central bank lowers the interest rate, more people will buy homes and cars."

But economics isn't physics. We don't fully understand the detailed causal processes, and we often can't do experiments (though in Unit 4 we will give examples of the use of conventional experiments in one area of economics). So how can economists do science? This example shows how the things we observe in the world can help us investigate causes and effects.

HOW ECONOMISTS LEARN FROM FACTS

DO INSTITUTIONS MATTER FOR GROWTH IN INCOME?

We can observe that capitalism emerged at the same time as, or just before, both the Industrial Revolution and the upward turn in our hockey sticks. This would be consistent with the hypothesis that capitalist institutions were among the causes of the era of continuous productivity growth. But the emergence of a free-thinking cultural environment known as the Enlightenment also predated or coincided with the upturn in the hockey sticks. So was it institutions, or culture, both, or some other set of causes? Economists and historians disagree, as you will see in Unit 2, when we ask "What were the causes of the Industrial Revolution?"

Scholars in all fields try to narrow the range of things on which they disagree by using facts. For complicated economic questions, like "Do institutions matter economically?", facts may provide enough information to reach a conclusion.

A method for doing this is called a *natural experiment*. It is a situation in which there are differences in something of interest—a change in institutions for example—that are not associated with differences in other possible causes. Because we cannot change the past, even if it were practical to conduct controlled experiments on entire populations, we rely on natural experiments, as Jared Diamond, a biologist, and James Robinson, a professor of government, explain.

The division of Germany at the end of the second world war into two separate economic systems—one centrally planned in the east, the other capitalist in the west—provides a natural experiment. The so-called Iron Curtain that divided them separated two populations sharing the same language, culture, and recent history as capitalist economies.

Before the second world war, living standards in what later became East and West Germany were the same. This is a suitable setting for using the natural experiment method. Before the war, firms in Saxony and Thuringia were world leaders in

automobile and aircraft production, chemicals, optical equipment and precision engineering.

With the introduction of centralised planning in East Germany, private property, markets and firms virtually disappeared. Decisions about what to produce, how much and in which plants, offices, mines and farms were taken not by private individuals, but by government officials. The state officials managing these economic organisations did not need to follow the principle of capitalism and produce goods and services that customers would buy at a price above their cost.

West Germany remained a capitalist economy.

The East German Communist Party forecast in 1958 that material wellbeing would exceed the level of West Germany by 1961. The failure of this prediction was one of the reasons the Berlin Wall separating East from West Germany was built in 1961. By the time the Berlin Wall fell in 1989, and East Germany abandoned central planning, its GDP per capita was less than half of that of capitalist West Germany. Figure 1.10 shows the different paths taken by these and two other economies from 1950. It uses the ratio scale.

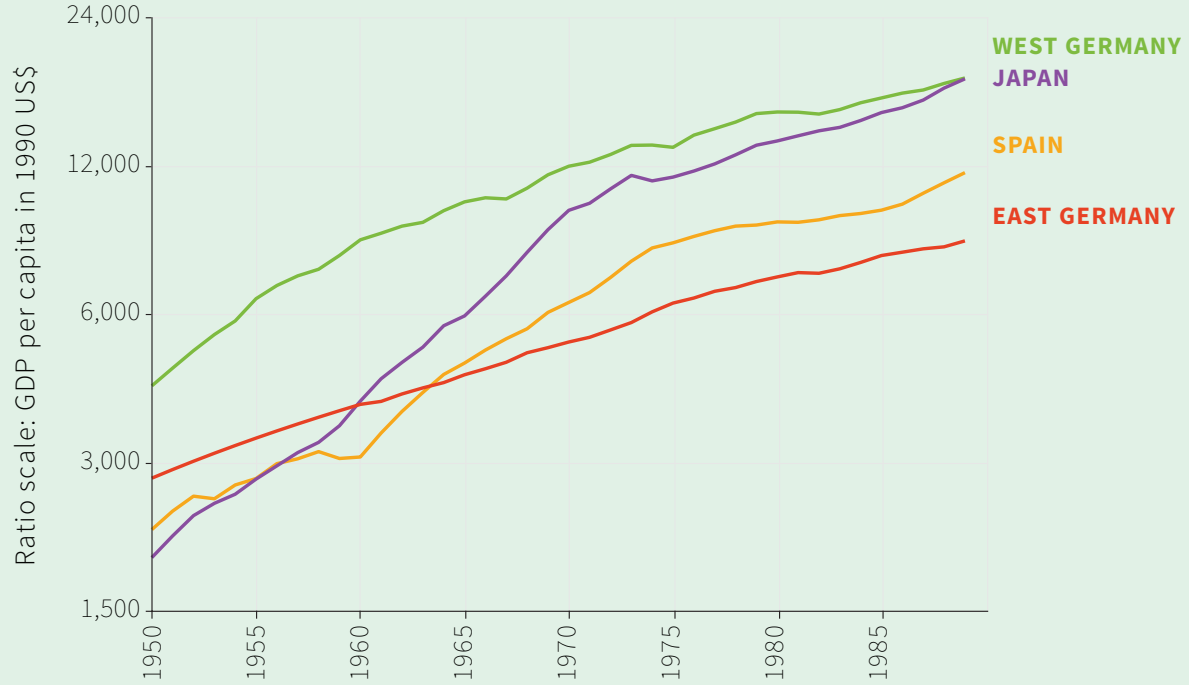


Figure 1.10 *The two Germanies: Planning and capitalism (1950-89).*

Source: *The Conference Board. 2015. 'Total Economy Database.' Accessed June 2015. Maddison, Angus. 2001. 'The World Economy: A Millennial Perspective.' Development Centre Studies. Paris: OECD.*

Notice from Figure 1.10 that East Germany had a less good starting position than West Germany in 1950. This was not mainly because of differences in the amount of capital equipment or skills available per head of the population, but because the structure of industries in East Germany was more disrupted by splitting the country than was the case in West Germany.

Unlike some capitalist economies that had lower per capita incomes in 1950, East Germany did not catch up to the world leaders, which included West Germany. By 1989, the Japanese economy (which had also suffered war damage) had, with its own particular combination of private property, markets and firms along with a strong government coordinating role, caught up to West Germany, and Spain had closed part of the gap.

We cannot conclude from the German natural experiment that capitalism always promotes rapid economic growth while central planning is a recipe for stagnation. Instead what we can infer is more limited: during the second half of the 20th century, the divergence of economic institutions mattered for the livelihoods of the German people.

When is capitalism dynamic?

Two sets of conditions contribute to the dynamism of the capitalist economic system. One set is economic; the other is political, and it concerns the government and the way it functions.

Economic conditions

The impact of economic conditions is summarised by the contrast between the second and third columns in Figure 1.11. Capitalism is less dynamic when property rights are insecure, there is limited competition in markets, and when the leadership of firms is in the hands of those who have not been tested by competition, but who instead have acquired their position via inheritance from parents or a political connection.

CHARACTERISTICS OF	WHEN CAPITALISM IS DYNAMIC	... AND WHEN IT ISN'T
PRIVATE PROPERTY	Secure	Insecure
MARKETS	Competitive (the losers lose)	Monopolised (the losers survive)
FIRMS	Leadership acquired by merit	Leadership from connections or inheritance

Figure 1.11 *Economic institutions that make capitalism dynamic.*

When these institutions are functioning well so that private property is secure, markets are competitive and firms led by people who have proven their merit, capitalism is unique. It is the first economic system in human history in which membership of the elite depends on a high level of economic performance.

As a firm owner, if you fail, you are no longer part of the club. Nobody kicks you out, because that is not necessary: you simply go bankrupt. An important feature of the discipline of the market—produce good products cheap or fail—is that it where it works well it is automatic; having a friend in power somewhere is no guarantee that you could remain in business. The same discipline applies to firms and to individuals in firms: losers lose. Market competition provides a mechanism for weeding out those who underperform.

Think of how different this is from other economic systems. A feudal lord who managed his estate poorly was just a shabby lord. But the owner of a firm that could not produce goods that people would buy, at prices that more than covered the cost, as we have seen is bankrupt—and a bankrupt owner is an ex-owner.

Of course, if they are initially very wealthy or very well connected politically, owners and managers of capitalist firms survive, and firms too stay in business despite their failures, sometimes for long periods or even over generations. Losers sometimes survive. But there are no guarantees: staying ahead of the competition means constantly innovating.

Political conditions

Government is also important. The policies it adopts often determine whether private property is secure, markets competitive, and firm leadership is based on merit. And these conditions determine how the carrots and sticks of the competitive process work.

For innovators to take the risk of introducing a new product or production process, their ownership of the resulting profits must be protected from theft by a well-functioning legal system. Governments also adjudicate disputes over ownership and enforce the property rights necessary for markets to work.

But, as Adam Smith warned, by creating monopolies such as the East India Company, governments may also take the teeth out of competition. If a large firm is able to establish a monopoly by excluding all competitors or a group of firms is able to collude to keep the price high, the incentives for innovation and the discipline of prospective failure will be dulled. The same is true in modern economies when some banks or other firms are considered to be *too big to fail* and instead are bailed out by governments when they might otherwise have failed.

In addition to providing an environment that supports the institutions of the capitalist economic system, the government provides essential goods and services such as physical infrastructure, education and national defence.

In a nutshell, capitalism can be a dynamic economic system when it combines:

- *Private incentives* for cost reducing innovation deriving from market competition and *secure private property*.
- Firms led by *those with proven ability* to produce goods at low cost.
- *Public policy* supporting these conditions, and supplying other essential goods and services.

These are the three conditions that together make up what we term the *capitalist revolution* that, first in Britain and then in some other economies, transformed the way that people interact with each other and with nature in producing their livelihoods.

1.11 VARIETIES OF CAPITALISM: DIVERGENCE AMONG LATECOMERS

Not every capitalist country is the kind of economic success story exemplified in Figure 1.1a by Britain, later Japan, and the other countries that caught up. Figure 1.12 tracks the fortunes of a selection of countries across the world during the 20th century. It shows for example that in Africa the success of Botswana in achieving sustained growth contrasts sharply with Nigeria's relative failure. Both are rich in natural resources (diamonds in Botswana, oil in Nigeria) and differences in the quality of their institutions—the extent of corruption and misdirection of government funds, for example—may help explain their contrasting trajectories.

The star performer in Figure 1.12 is South Korea. In 1950 its GDP per capita was the same as Nigeria's; in 2013 it was 10 times richer by this measure. South Korea's takeoff occurred under institutions and policies sharply different from those prevailing in Britain in the 18th and 19th centuries. The most important difference is that the government of South Korea (along with a few very large corporations) played a leading role in directing the process of development, explicitly promoting some industries, requiring firms to compete in foreign markets and also providing high quality education for its workforce. The term *developmental state* has been applied to the leading role of the South Korean government in its economic takeoff and now refers to any government playing this part in the economy. Japan and China are other examples of developmental states.

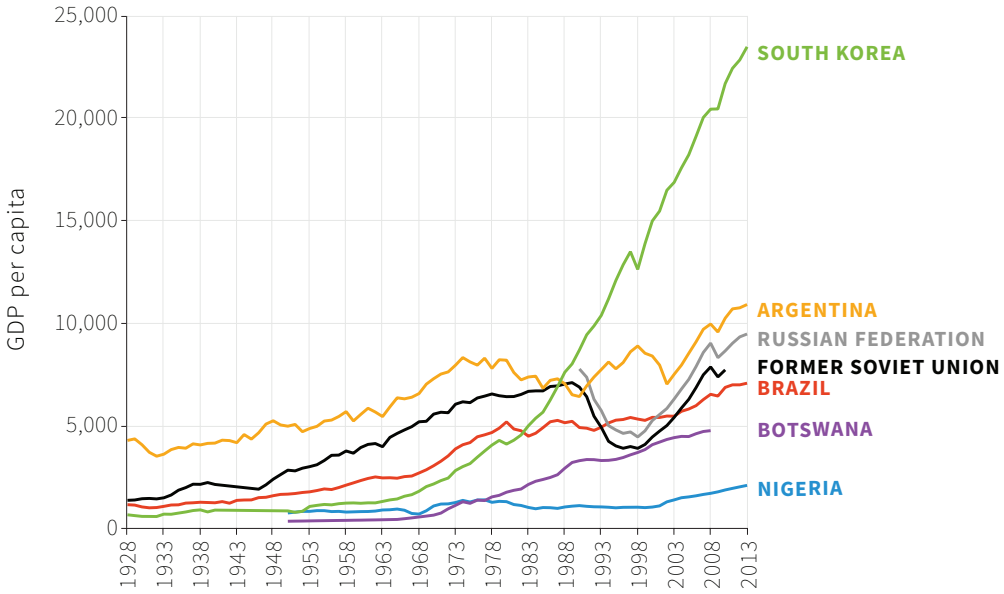


Figure 1.12 *Divergence of GDP per capita among latecomers to the capitalist revolution (1928-2013).*

Source: *Bolt, Jutta, and Jan Juiten van Zanden. 2013. 'The First Update of the Maddison Project Re-Estimating Growth Before 1820.' Maddison-Project Working Paper WP-4, January.*

From Figure 1.12 we also see that in 1928, when the Soviet Union's first five-year economic plan was introduced, GDP per capita was one-tenth of the level in Argentina, similar to Brazil, and considerably higher than in South Korea. Central planning in the Soviet Union produced steady but unspectacular growth for nearly 50 years. GDP per capita in the Soviet Union outstripped Brazil by a wide margin and even overtook Argentina briefly just before Communist party rule there ended in 1990.

The contrast between West and East Germany demonstrates that one reason central planning was abandoned as an economic system was its failure, in the last quarter of the 20th century, to deliver the improvements in living standards achieved by some capitalist economies. Yet the varieties of capitalism that replaced central planning in the countries that had once made up the Soviet Union did not work so well either. This is evident from the pronounced dip in GDP per capita for the ex-Soviet Union after 1990, shown in Figure 1.12.

The lagging performances of some capitalist economies, including the ones in Figure 1.12 in which growth was slow or uneven, highlight the following problems from the right-hand column of Figure 1.11:

- *Private property may not be secure* as a result of weak enforcement of the rule of law and of contracts, or expropriation either by criminal elements or by government bodies.
- *Markets may not be competitive* and may fail to offer the carrots and wield the sticks that make a capitalist economy dynamic.

- Partly as a result of these failures, *firms may be owned and managed by people who survive because of their connections to government or their privileged birth* rather than their aptitude for delivering high quality goods and services at a competitive price.

Combinations of failures of the three basic institutions of capitalism mean that individuals and groups often have more to gain by spending time and resources in lobbying, criminal activity, and other ways of seeking to shift the distribution of income in their favour, and less in the creation of wealth.

1.12 VARIETIES OF CAPITALISM: GOVERNMENT AND THE ECONOMY

We have seen that in some economies—South Korea for example—governments have played a leading role in the capitalist revolution. But even where government's role is more limited, as in Britain at the time of its takeoff, governments establish, enforce and change the laws and regulations that influence how the economy works. For example, markets, private property and firms are all regulated by laws and policies. Moreover, in virtually every modern capitalist economy, governments are a large part of the economy, accounting in some for more than half of the economy's GDP.

In subsequent units we investigate why government policies in such areas as sustaining competition, taxing and subsidising to protect the environment, influencing the distribution of income, the creation of wealth, and the level of employment and inflation may make good economic sense.

One of the reasons why capitalism comes in so many forms is that over the course of history and today, capitalist economies have coexisted with many political systems. A *political system* such as *democracy* determines how governments will be selected, and how those governments will make and implement decisions that affect the population.

Capitalism emerged in Britain, the Netherlands, and in most of today's high-income countries long before democracy. In no country were most adults eligible to vote prior to the end of the 19th century (New Zealand was the first). Even in the recent past, capitalism has coexisted with undemocratic forms of rule, as in Chile from 1973-90, in Brazil from 1964-85, and in Japan until 1945. Contemporary China has a variant of the capitalist economic system, but its system of government is not a democracy by our definition. In most countries today, however, capitalism and democracy coexist, each system influencing how the other works.

Like capitalism, democracy comes in many forms. In some, the head of state is elected directly by the voters; in others it is an elected body, such as a parliament, that elects the head of state. In some democracies there are strict limits on the ways in which individuals can influence elections or public policy through their financial contributions; in others private money has great influence through contributions to electoral campaigns, lobbying, and even illicit contributions, such as bribery.

DEMOCRACY

Democracy is one among many political systems, defined by:

- Individual rights including freedom of speech, assembly, and the press
- Fair elections in which virtually all adults are eligible to vote
- ... and in which the loser leaves office

These differences even among democracies are part of the explanation of why the government's importance in the capitalist economy differs so much among nations. In Figure 1.13 we show one measure of the size of government relative to the entire economy: the total amount of taxes collected by government (both local and national) as a fraction of GDP. Even among economies at about the same level of GDP per capita, the size of government by this measure varies. In the US it is one-third; in six rich countries in northern Europe, it is more than a half.

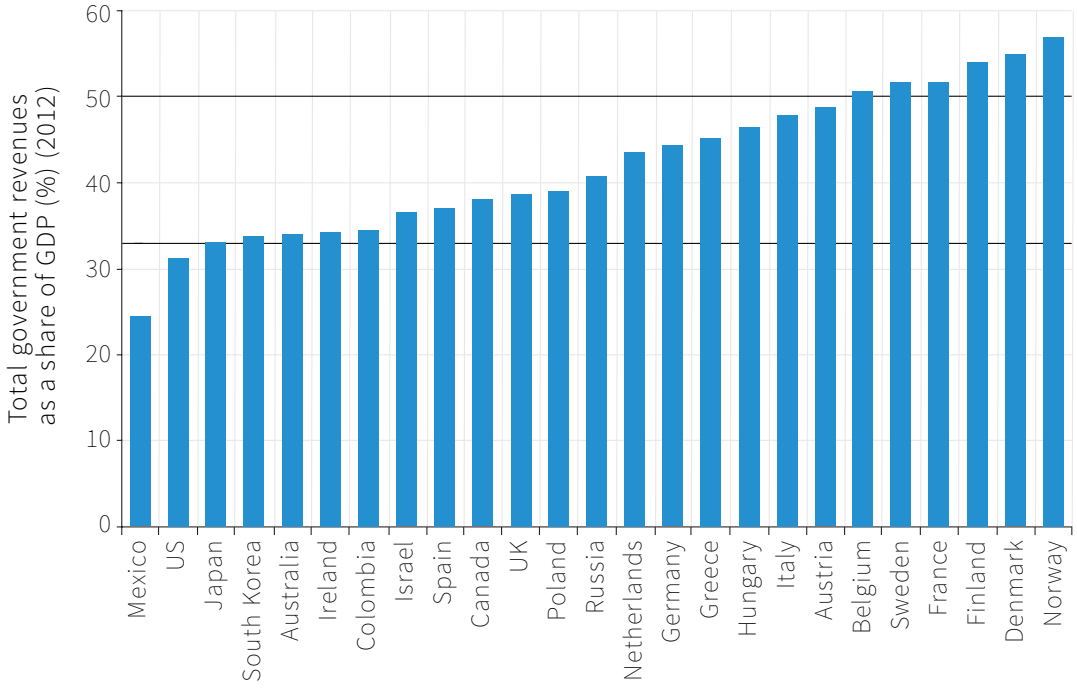


Figure 1.13 *The size of government as measured by total tax revenue as a fraction of GDP (2012).*

Source: *OECD (2015), General government revenue indicator.*

Notice that by this measure the developmental state South Korea resembles the US, although in the US the government takes a much less active role in directing the economy. Government revenues are limited in Japan, too. But the governments of Japan and South Korea play an important role in setting the direction of their economies, just as important as the role of governments of Sweden and Denmark that spend a far greater fraction of the total income of the country.

The big difference between South Korea and Japan on the one hand, and Sweden and Denmark on the other, is the extent to which government policies reduce the amount of inequality in disposable income. We will see in the next section that in Sweden and Denmark inequality in disposable income (by one of the most commonly used measures) is just half the level of inequality of income before the payment of taxes and receipt of transfers. In Japan and South Korea, by contrast, government taxes and transfers also reduce inequality in disposable income, but to a far lesser degree.

1.13 MEASURING ECONOMIC INEQUALITY

The measure of inequality that we referred to in the comparison of government policies in Japan, South Korea, Sweden and Denmark is called the *Gini coefficient* after the Italian statistician Corrado Gini (1884-1965). It indicates how much disparity there is in income, or any other measure, across the population. If everyone has the same income, so there is no inequality, the Gini coefficient takes a value of 0. The maximum inequality, a value of 1, means a single individual receives all the income.

When we pointed out that government taxes and transfers in Sweden created a distribution of income that was half as unequal as before taxes and transfers, we were saying that Sweden's Gini coefficient *before taxes and transfers* (the equivalent for the Netherlands is shown in Figure 1.14a below) is 0.47, while the Gini coefficient for disposable income in Sweden is 0.24.

Like GDP per capita, the Gini coefficient measures an important characteristic about the whole economy. And, like GDP per capita, it is worth exploring exactly what the Gini coefficient measures.

The Gini coefficient is based on a statistical construct called the *Lorenz curve* (invented in 1905 by Max Lorenz (1876-1959), an American economist, while he was still a student). We will explain the Lorenz curve before showing how you calculate the Gini coefficient from it.

The Lorenz curve shows the entire population lined up along the horizontal axis from the poorest to the richest. The height of the curve at any point on the horizontal axis indicates the fraction of total income received by the fraction of the population given by that point on the horizontal axis.

Figure 1.14a shows a Lorenz curve in the Netherlands in 2010. It is based on data for market income so it does not take account of taxes and government transfers (we will see what difference they make soon). The curve indicates that the poorest 10% of the population (10 on the horizontal axis) receives only 0.1% of the total income (0.1 on the vertical axis). The other points in the curve convey the same kind of information.

When studying large populations like that of a country or city, as is usually the case, the Gini coefficient is the area between the perfect equality line and the Lorenz curve (denoted *A* in Figure 1.14a), divided by the entire area under the perfect equality line (*A* + *B*). The Gini coefficient was introduced by the Italian statistician just seven years after Lorenz came up with his curve. So:

$$Gini = \frac{A}{A+B}$$

From the data we used to construct the Lorenz curve we calculate that the Gini coefficient of market income in the Netherlands in 2010 is 0.47.

As you can see from the slideline, when we draw the Lorenz curve for disposable income, the new shaded area *A'* is much smaller, and the new Gini correspondingly lower:

$$Gini = \frac{A'}{A'+B'}$$
$$= 0.25$$

This shows us that in the Netherlands, as in Sweden and Denmark, taxes and transfers substantially reduce disparities in disposable income.

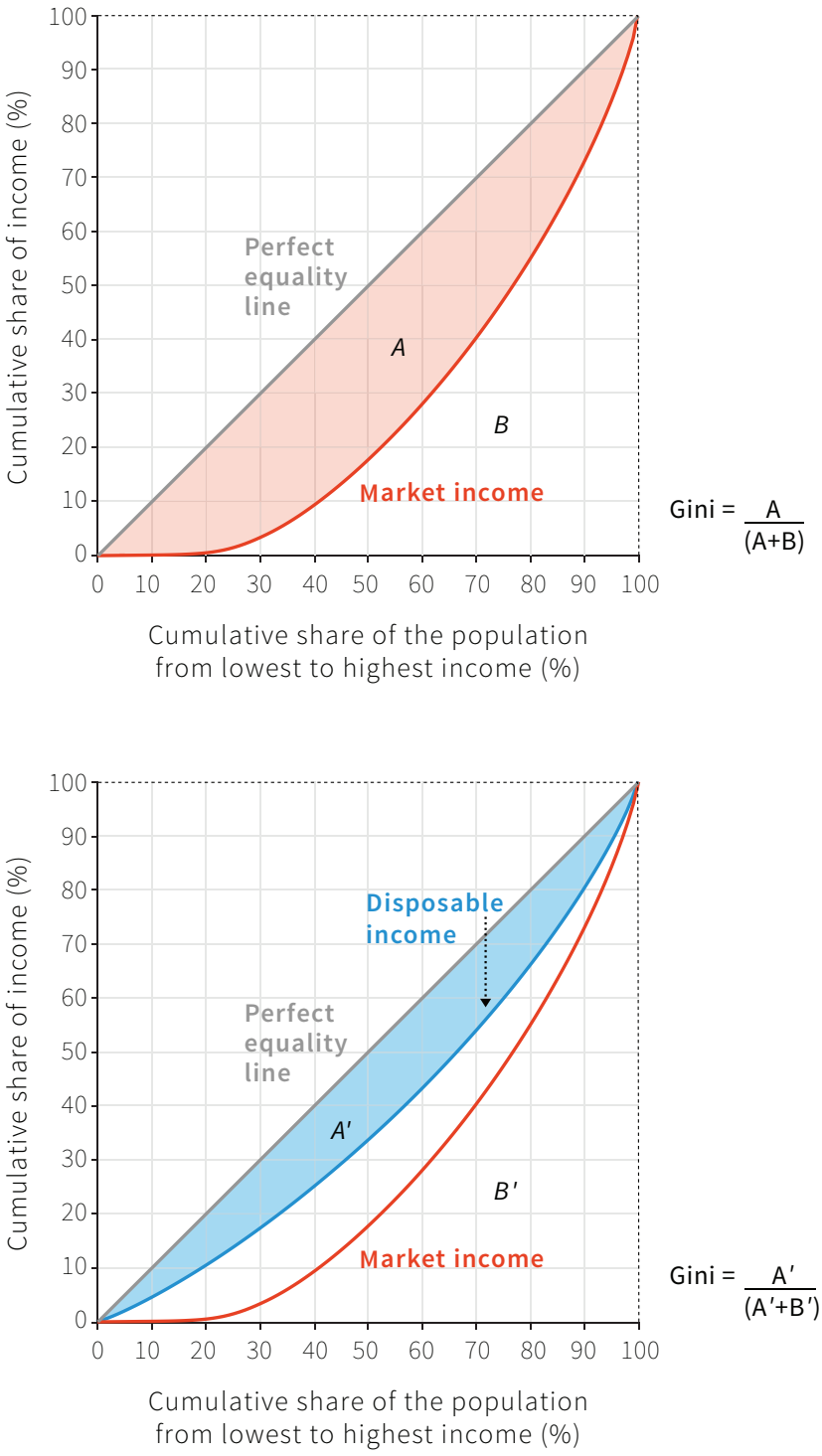


Figure 1.14a *Distribution of market and disposable income in the Netherlands (2010).*

Source: *LIS. 2015. 'Cross National Data Center.' LIS. Accessed June. Calculations were conducted by Stefan Thewissen (University of Oxford) in April 2015. Household market (labour and capital) income and disposable income are equivalised and top- and bottom-coded.*

Income inequality in the Netherlands

Notice that the Lorenz curve is bowed downwards from the 45-degree line in the figure. This is because there are income inequalities among people in the Netherlands. The 45-degree line is what the Lorenz curve would look like if everyone had the same income. Because it has a slope of 1 the poorest 10% receives 10% of income, and so on. In this case, we would not have to line people up according to their income: they would all be at the front of the queue. The shaded area labelled A shows how far the Lorenz curve is bowed out from the 45-degree line of equality. This is a measure of the extent of income inequality in the Netherlands. Now compare the Lorenz curve for disposable income with the curve for market income. The new shaded area A′ between the disposable income curve and the perfect equality line is much smaller. This is because taxes and transfers have reduced inequality in disposable income.

Like GDP per capita, we can also use the Lorenz curve and the Gini coefficient to compare countries. For example Figure 1.14b shows the Lorenz curve for disposable income in the US. Comparing this to the analogous curve for disposable income for the Netherlands we see that the US is much more unequal by this measure: using again the formula for the Gini, we find that the Gini coefficient for disposable income in the US is 0.39.

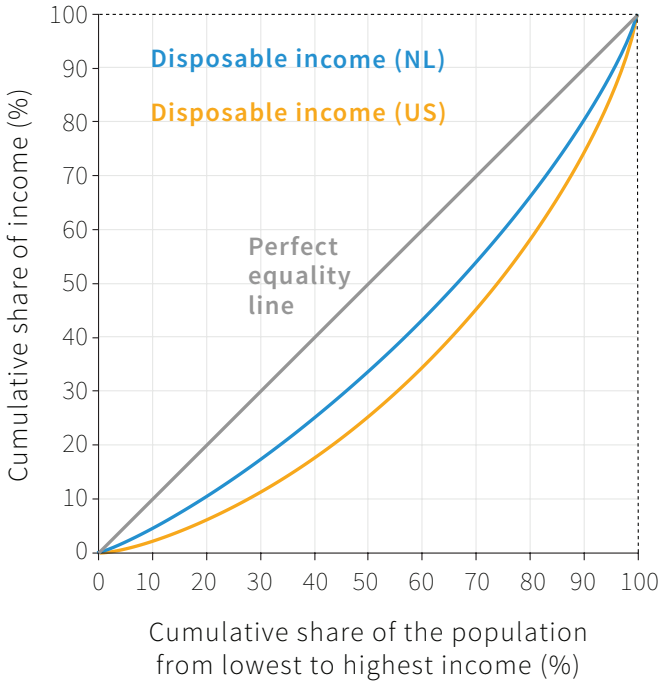


Figure 1.14b *Distribution of disposable income in the Netherlands (2010) and the US (2013).*

Source: *LIS. 2015. 'Cross National Data Center.' LIS. Accessed June.*

These just two of many ways to measure inequality. Others you may see include the fraction of all income received by the richest 1% of the population, or the ratio of the income at the 90th percentile of income to the income at the 10th percentile.

1.14 VARIETIES OF CAPITALISM: ECONOMIC INEQUALITY

The Gini coefficient (or alternative measures such as the share of income received by the top income recipients) can, like GDP per capita, be used to track trends in a given country over time.

Gini coefficients for income since the 18th century in the US, Britain and the Netherlands are shown in Figure 1.15. There has been a more or less continuous decline in income inequality in the Netherlands since the middle of the 18th century. In Britain inequality rose during the late 18th century, and then fell until the closing decades of the 20th century, after which it increased again. In the US, inequality rose from the time of the Declaration of Independence in 1776 until the Civil War in 1861, and then declined for the next century, only to rise again in recent years. Inequality of income in the US, as measured by the Gini coefficient, is now slightly higher than it was when slavery existed, on the eve of the American Civil War.

The sharp increase in inequality in Britain and the US in recent years has also occurred in some major economies, such as India and China, but not in others.

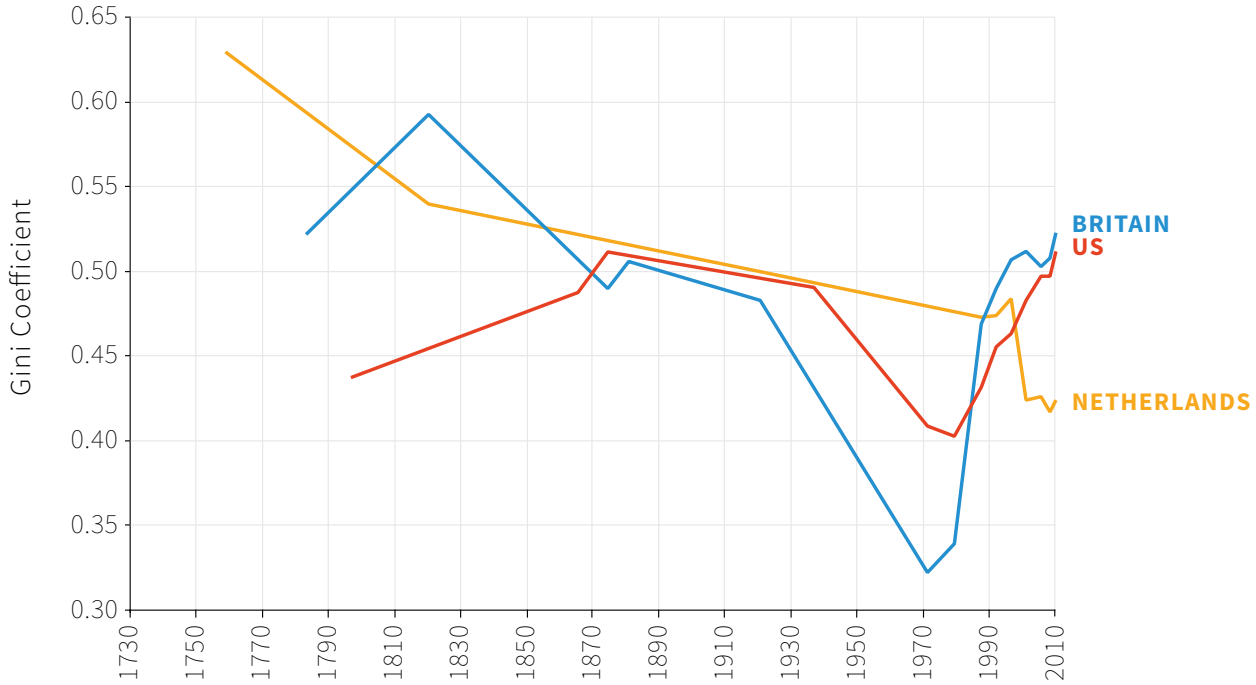


Figure 1.15 *Income Inequality in the US, Britain and the Netherlands (1730-2010).*

Source: *Lindert, Peter, and Jeffrey Williamson. 2103. 'Two Centuries of American Growth and Inequality, 1650-1860.' Stanford Economic History Seminar, October. The figure measures inequality of market, not disposable income for which data are not available before recent years, so the effects of taxes and transfers are not included. But prior to 1950 these were of limited importance.*

Figure 1.15 (and the comparison of the US and the Netherlands in the previous section) illustrate two important points about capitalism and inequality:

- *Change over time*: A capitalist economy may become less unequal over time, or more unequal.
- *Differences between economies*: At a fixed point in time the degree of inequality in disposable income may differ dramatically between different capitalist economies, with some highly unequal and others much less so.

The main reason for the differences between nations in inequality of disposable income is the extent to which governments tax wealthy families and transfer the proceeds to less well off individuals. Figure 1.16 shows inequality of both market and disposable income as measured by the Gini coefficient. The top of the lower part of each bar gives the Gini for disposable income; the top of the upper part of the bar shows the Gini for market income. The countries are ordered from left to right, from the least to the most unequal by the disposable income measure (because this is the preferred measure of inequality in living standards).

Notice that:

- The differences between countries in inequality in disposable incomes (the top of the lower bars) are much greater than inequalities in income before taxes and transfers (the top of the upper bars).
- The US and the UK are among the most unequal of the high-income economies.
- The few poor and middle income countries for which data are available are even more unequal in disposable income than the US.

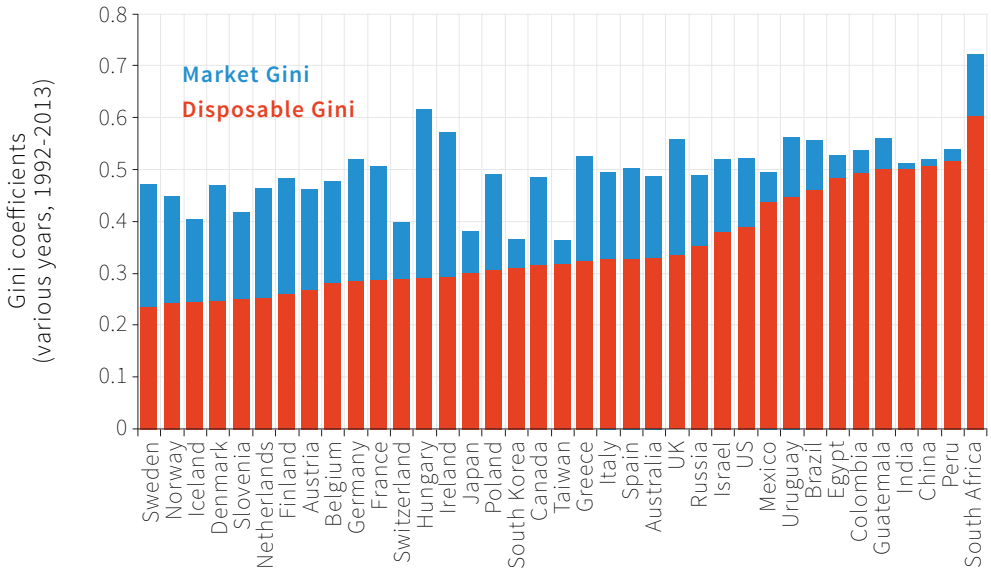


Figure 1.16 *Income inequality in market and disposable income across the world.*

Source: *LIS. 2015. 'Cross National Data Center.' LIS. Accessed June. Estimates by Stefan Thewissen (University of Oxford) in April 2015. Household market (labour and capital) income and disposable income are equivalised and top- and bottom-coded.*

But (with the exception of South Africa) this is mainly the result of the very limited degree of redistribution from rich to poor, not an unusually high degree of inequality in income before taxes and transfers.

Figure 1.17 shows—for the same countries as in Figure 1.16—a measure of the extent to which taxes and transfers distribute income to the less well off. This is the redistribution ratio, namely the length of the blue segment in Figure 1.16 divided by the total height of the bar (top of the blue bar).

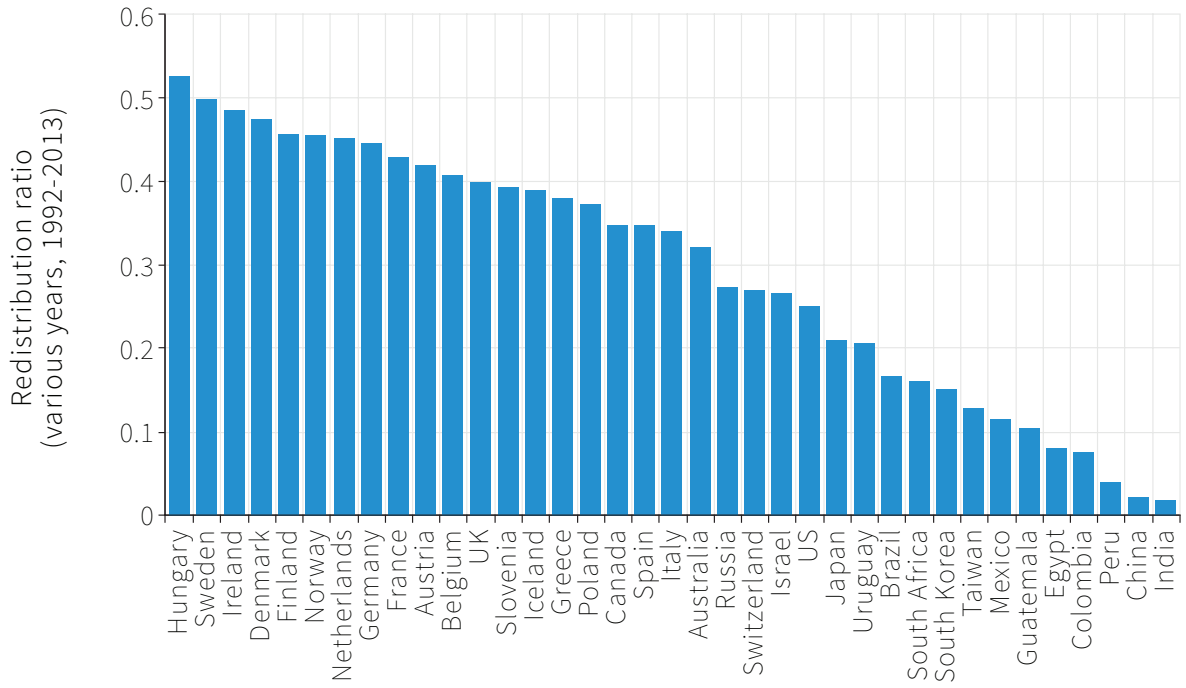


Figure 1.17 *Redistribution ratios across the world.*

Source: *LIS. 2015. 'Cross National Data Center.' LIS. Accessed June. Estimates by Stefan Thewissen (University of Oxford) in April 2015. Household market (labour and capital) income and disposable income are equivalised and top- and bottom-coded.*

DISCUSS 1.8: THE REDISTRIBUTION RATIO

From Figure 1.17, select two countries that have very different redistribution ratios.

Referring to the politics, history and economics of these countries, explain why these ratios might be so different.

1.15 ECONOMICS AND THE ECONOMY

Economics is the study of how people interact with each other and with their natural surroundings in producing their livelihoods, and how this changes over time. Therefore it is about:

- *How we come to acquire the things*—food, clothing, shelter, free time—that make up our livelihood and, in doing this,
- *How we interact with each other* either as buyers and sellers, employees or employers, citizens and public officials, parents, children and other family members.
- *How we interact with our natural environment*, from breathing to extracting raw materials from the earth.
- *How each of these changes* over time.

In Figure 1.8 we showed that the economy is part of society, which in turn is part of the biosphere. Figure 1.18 shows the position of firms and families in the economy, and the flows that occur within the economy and between the economy and the biosphere. Firms combine labour with structures and equipment, and produce goods and services that are used by households and other firms.

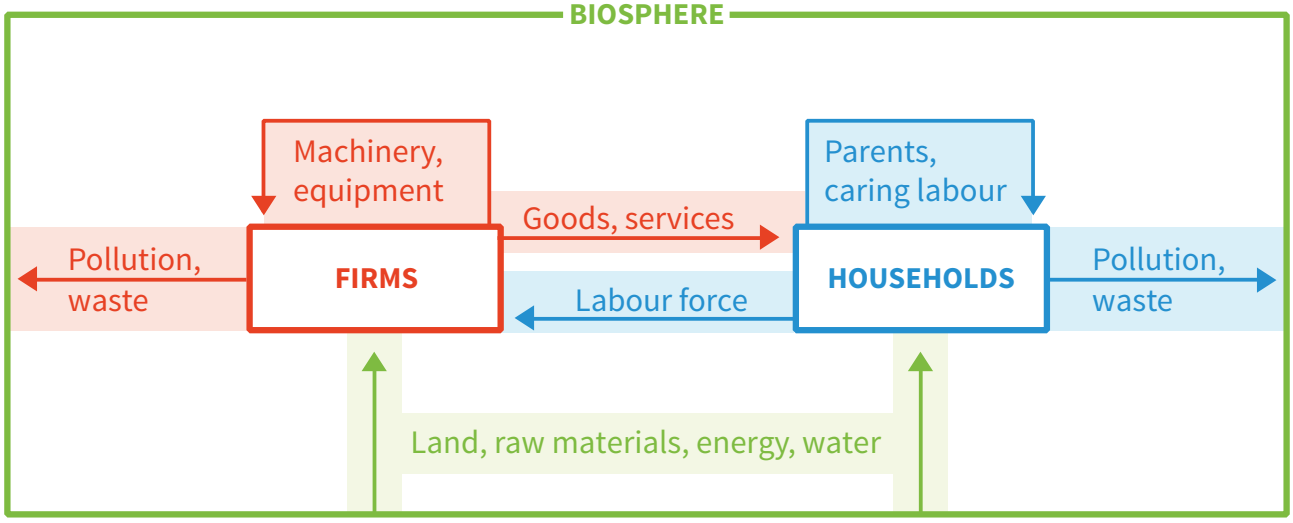


Figure 1.18 *A model of the economy: Households and firms.*

Production of goods and services also takes place within households although, unlike firms, households may not sell their outputs in the market. In addition to producing goods and services, households are also producing people—the next generation of the labour force. The labour of parents, care givers and others is combined with

structures (for example, your home) and equipment (for example, the oven in that home) to reproduce and raise the future labour force working in firms, and the people who will work and reproduce in the households of the future.

All of this takes place as part of a biological and physical system in which both firms and households make use of our natural surroundings and resources, ranging from fossil fuel based energy to the air we breathe. In the process, households and firms transform nature by using its resources, but also by producing inputs to nature. Currently some of the most important of these inputs are the greenhouse gases, which contribute to the climate change problems that we saw in section 1.6.

1.16 CONCLUSION

Capitalism is the most dynamic economic system the world has ever known. So far, this has been mostly good news: many capitalist economies have brought substantial, sustained increases in access to material goods and to free time for their citizens.

On the other hand, despite the permanent technological revolution, material deprivation and insecurity persist, and many people consider the extent of income disparities among households unfair.

While capitalism's dynamism has the potential to create technologies that will lessen pollution, innovation that is unregulated by environmental policy poses a threat to the natural surroundings on which life depends.

CONCEPTS INTRODUCED IN UNIT 1

Before you move on, review these definitions:

- Economics
- Industrial Revolution
- Technology
- Economic system
- Capitalism
- Institutions
- Private property
- Markets
- Firms
- Capitalist revolution
- Democracy

DISCUSS 1.9: WHERE AND WHEN WOULD YOU CHOOSE TO HAVE BEEN BORN?

Suppose you can choose to be born in any time period in any of the countries in Figure 1.1a, 1.10 or 1.12, but you know that you would be among the poorest 10% in the population.

1. In which country would you choose to be born?
2. Suppose, instead, you know you would be among the poorest 10% in the population, but you can move to the top 10% of the population if you work hard. In which country would you now choose to be born?
3. Finally, suppose that you can only decide on the country and time period of your birth. You cannot be sure if you would be born in the city or the countryside, would be male or female, rich or poor. In which time and country would you choose to be born?
4. For the scenario in (3), in which time and country you would least want to be born?

Use what you have learned from this unit to explain your choices.

Key points in Unit 1

GDP and GDP per capita

Gross domestic product is a measure of the income of a country. GDP per capita is GDP divided by population, and is commonly used as a measure of living standards.

The hockey stick

Throughout most of history GDP per capita was relatively similar around the world, and it changed little from century to century. Since 1700 it has risen rapidly in some countries, led by Britain.

The permanent technological revolution

The period since 1700 has also seen improvements in technology, increases in population, impacts on the environment and differences in income among countries.

Capitalism

Capitalism is an economic system in which firms, private property and markets play a major role.

Impacts of capitalism

Along with the permanent technological revolution this new economic system has revolutionised the way people interact with each other, and with nature, in producing their livelihoods.

Inequality

Inequality among a group of people is measured by the Lorenz curve and the Gini coefficient.

Divergence

Capitalist economies throughout the world, and in the past, differ greatly in the form of governments and public policies, the degree of inequality, and the extent of improvements in living standards.

1.17 EINSTEIN

Comparing income at different times, and across different countries

The United Nations brings together estimates of GDP from statistical agencies around the world. These estimates, with those made by economic historians, allow us to construct charts like Figure 1.1a, comparing living standards across countries and at different time periods, and looking at whether the gap between rich and poor countries has narrowed or widened over time. Before we can make a statement like: "On average, people in Italy are richer than people in China, but the gap between them is narrowing," statisticians and economists must try to solve three problems:

- We need to separate the thing we want to measure—changes or differences in amounts of goods and services—from things that are not relevant to the comparison, especially changes or differences in the prices of the goods and services.
- When comparing output *in one country at two points in time*, it is necessary to take into account differences in prices between the two points in time.
- When comparing output *between two countries at a point in time*, it is necessary to take into account differences in prices between the two countries.

Notice how similar the last two statements are. Measuring changes in output at different points in time presents the same challenges as we face when we try to compare countries by measuring differences in their output at the same time. The challenge is to find a set of prices to use in this calculation that will allow us to identify changes or differences in outputs, without making the mistake of assuming that if the price of something rises in a country, but not in another, then the amount of output has increased in the country.

The starting point: Nominal GDP

When estimating the market value of output in the economy as a whole for a given period, such as a year, statisticians use the prices at which goods and services are sold in the market. By multiplying the quantities of the vast array of different goods and services by their prices, they can be converted into money, or nominal, terms. With everything in the common unit of nominal (or money) terms, they can be added together. *Nominal GDP* is written like this:

(price of a yoga lesson × number of yoga lessons) + (price of a book × number of books) + . . . + (price x quantity) for all other goods and services

In general, we write that:

$$nominal\ GDP \equiv \sum_i p_i q_i$$

where p_i is the price of good i, q_i is the quantity of good i, and $\sum$ indicates the sum of price time quantity for all the goods and services that we count.

Taking account of price changes over time: Real GDP

To gauge whether the economy is growing or shrinking, we need a measure of the *quantity* of goods and services purchased. This is called *real GDP*. If we compare the economy in two different years, and if all the quantities stay the same, but the prices increase by, say, 2% from one year to the next, then nominal GDP rises by 2%, but real GDP is unchanged. The economy has not grown.

Because we cannot add together the number of computers, shoes, restaurant meals, flights, fork-lift trucks and so on, it is not possible to measure real GDP directly. Instead, to get an estimate of real GDP, we have to begin with nominal GDP as defined above.

On the right-hand side of the equation for nominal GDP are the prices of each item of final sales multiplied by the quantity.

To track what is happening to real GDP, we begin by selecting a base year: for example, the year 2010. We then define real GDP using 2010 prices as equal to nominal GDP that year. The following year, nominal GDP for 2011 is calculated as usual using the prices prevailing in 2011. Next, we can see what has happened to real GDP by multiplying the 2011 quantities by the 2010 prices. If, using the base year prices, GDP has gone up, we can infer that real GDP has increased.

If this method produces the result that, when computed using 2010 prices, GDP in 2011 is the same as in 2010, we can infer that although there might have been a change in the composition of output (fewer flights taken but more computers sold, for example), the overall quantity of output of goods and services has not changed. The conclusion would be that real GDP, which is also called GDP at *constant prices*, is unchanged. The growth rate of the economy *in real terms* is zero.

Taking account of price differences among countries: International prices and purchasing power

To compare countries, we need to choose a set of prices and apply that set of prices to both countries.

To begin with, imagine a simple economy which produces only one product. In our example, we choose a regular cappuccino because we can easily find out the price of this standard product in different parts of the world. And we choose two economies that are very different in their level of development: Sweden and Indonesia.

When prices are converted into US dollars using current exchange rates, a regular cappuccino costs $3.76 in Stockholm and $2.71 in Jakarta. But simply expressing the two cappuccinos in a common currency is not enough, because the international

current exchange rate that we used to get these numbers is not a very good measure of how much a rupiah will buy in Jakarta and how much a krona will get you in Stockholm.

This is why when comparing living standards across countries we use estimates of GDP per capita in a common set of prices known as *Purchasing Power Parity (PPP)* prices. As the name suggests, the idea is to achieve parity (equality) in the real purchasing power.

Prices are typically higher in richer countries—as in our example. One reason for this is that wages are higher, which translates into higher prices. Because prices of cappuccinos, restaurant meals, haircuts, most types of food, transport, rents and most other goods and services are more expensive in Sweden than in Indonesia, once a common set of prices is applied, the difference between GDP per capita in Sweden and Indonesia measured at PPP is smaller than it is if the comparison is made at current exchange rates.

At current exchange rates, GDP per capita in Indonesia is only 6% the level of Sweden; at PPP where the comparison uses international prices, GDP per capita in Indonesia is 21% the level of Sweden.

What this comparison shows is that the buying power of the Indonesian rupiah compared to the Swedish krona is more than three times greater than would be indicated by the current exchange rate between the two currencies.

We will examine the measurement of GDP (and other measures of the whole economy) in more detail in Unit 12.

1.18 READ MORE

Bibliography

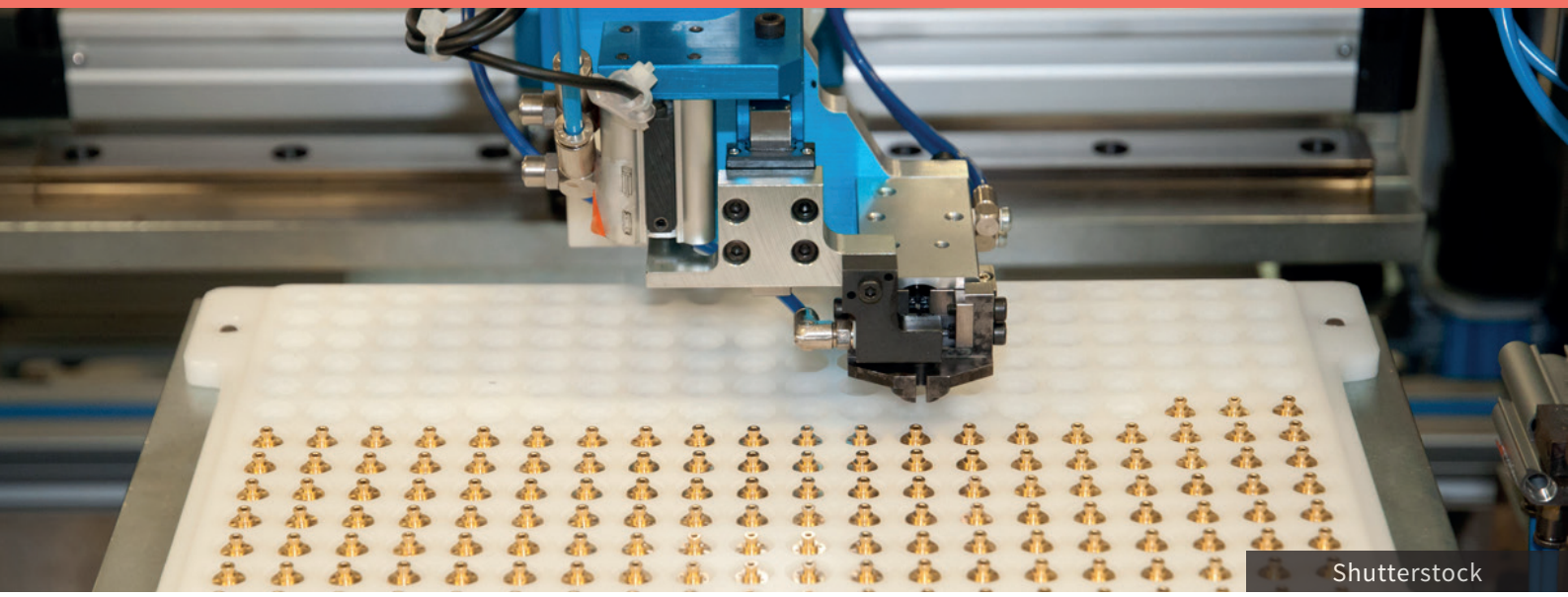
1. Acemoglu, Daron, and James A. Robinson. 2012. *Why Nations Fail: The Origins of Power, Prosperity, and Poverty*. New York, NY: Crown Publishing Group.
2. Amsden, Alice H. 1989. *Asia's next Giant: South Korea and Late Industrialization*. New York, NY: Oxford University Press.
3. Augustine, Dolores. 2013. 'Innovation and Ideology: Werner Hartmann and the Failure of the East German Electronics Industry.' In *The East German Economy, 1945-2010: Falling behind or Catching Up?*, by German Historical Institute, edited by Hartmut Berghoff and Uta Andrea Balbier. Cambridge: Cambridge University Press.
4. Berg, Maxine, and Pat Hudson. 1992. 'Rehabilitating the Industrial Revolution.' *The Economic History Review* 45 (1).
5. Berghoff, Hartmut, and Uta Andrea Balbier. 2013. 'From Centrally Planned Economy to Capitalist Avant-Garde? The Creation, Collapse, and Transformation of a Socialist Economy.' In *The East German Economy, 1945-2010 Falling behind or Catching Up?*, by German Historical Institute, edited by Hartmut Berghoff and Uta Andrea Balbier. Cambridge: Cambridge University Press.
6. Bolt, Jutta, and Jan Juiten van Zanden. 2013. 'The First Update of the Maddison Project Re-Estimating Growth Before 1820.' *Maddison-Project Working Paper WP-4*, January.
7. Bolt, Jutta, and Jan Luiten van Zanden. 2014. 'The Maddison Project: Collaborative Research on Historical National Accounts.' *The Economic History Review*.
8. Broadberry, Stephen. 2013. 'Accounting for the Great Divergence.' London School of Economics and Political Science. November 1.
9. Chang, Ha-Joon. 2008. *Bad Samaritans: The Guilty Secrets of Rich Nations and the Threat to Global Prosperity*. London: Cornerstone.
10. Cipolla, Carlo M. 1978. *The Economic History of World Population*. New York, NY: Barnes and Noble.
11. Clark, Andrew. 2015. 'Attitudes to Income Inequality: Experimental and Survey Evidence.' In *Handbook of Income Distribution*, edited by Antonio D'Ambrosio. Amsterdam, Oxford: Elsevier.
12. Clark, Andrew, and Andrew Oswald. 2002. 'A Simple Statistical Method for Measuring How Life Events Affect Happiness.'*International Journal of Epidemiology* 31 (6): 1139–44.
13. Clark, Gregory. 2007. *A Farewell to Alms: A Brief Economic History of the World*. Princeton, NJ: Princeton University Press.
14. Coyle, Diane. 2014. *GDP: A Brief but Affectionate History*. Princeton, NJ: Princeton University Press.

15. Crafts, Nicholas. 2004. 'Steam as a General Purpose Technology: A Growth Accounting Perspective*.' *The Economic Journal* 114 (495): 338–51.
16. Cronon, William. 2003. *Changes in the Land: Indians, Colonists, and the Ecology of New England*. 1st ed. New York, NY: Farrar, Straus and Giroux.
17. Deaton, Angus. 2013. *The Great Escape: Health, Wealth, and the Origins of Inequality*. Princeton, NJ: Princeton University Press.
18. Diamond, Jared. 1999. *Guns, Germs, and Steel: The Fates of Human Societies*. New York, NY: Norton, W. W. & Company.
19. Diamond, Jared, and James Robinson. 2014. *Natural Experiments of History*. Cambridge, MA: Belknap Press of Harvard University Press.
20. Dobb, Maurice. 1964. *Studies in the Development of Capitalism*. New York, NY: International Publishers.
21. Eurostat. 2015. 'Quality of Life Indicators - Measuring Quality of Life.' June.
22. Flannery, Tim F. 2002. *The Future Eaters: An Ecological History of the Australasian Lands and People*. 1st ed. New York, NY: Grove Press/Atlantic Monthly Press.
23. Friedman, Milton. 1962. *Capitalism and Freedom*. Chicago, IL: University of Chicago Press.
24. Kahneman, Daniel, and Angus Deaton. 2010. 'High Income Improves Evaluation of Life but Not Emotional Well-Being.' *Proceedings of the National Academy of Sciences* 107 (38): 16489–93.
25. Kornai, János. 2013. *Dynamism, Rivalry, and the Surplus Economy: Two Essays on the Nature of Capitalism*. Oxford: Oxford University Press.
26. Landes, David S. 1999. *The Wealth and Poverty of Nations: Why Some Are So Rich and Some Are So Poor*. New York, NY: Norton, W. W. & Company.
27. Landes, David S. 2003. *The Unbound Prometheus: Technological Change and Industrial Development in Western Europe from 1750 to the Present*. Cambridge: Cambridge University Press.
28. Lindert, Peter, and Jeffrey Williamson. 2103. 'Two Centuries of American Growth and Inequality, 1650-1860.' Stanford Economic History Seminar, October.
29. Lorenz, Max O. 1905. 'Methods of Measuring the Concentration of Wealth.' *Publications of the American Statistical Association* 9 (70).
30. Maddison, Angus. 2001. 'The World Economy: A Millennial Perspective.' *Development Centre Studies*. Paris: OECD.
31. Mann, M. E., Z. Zhang, M. K. Hughes, R. S. Bradley, S. K. Miller, S. Rutherford, and F. Ni. 2008. 'Proxy-Based Reconstructions of Hemispheric and Global Surface Temperature Variations over the Past Two Millennia.' *Proceedings of the National Academy of Sciences* 105 (36): 13252–57.
32. McNeill, John Robert R. 2000. *Something New under the Sun: An Environmental History of the Twentieth-Century World*. 1st ed. New York, NY: W.W. Norton & Co.
33. McNeill, John Robert R., and William H. McNeill. 2003. *The Human Web: A Bird's-Eye View of World History*. 1st ed. New York, NY: WW Norton & Co.

34. Nordhaus, William. 1998. 'Do Real Output and Real Wage Measures Capture Reality? The History of Lighting Suggests Not.' *Cowles Foundation For Research in Economics* Paper 957.
35. Piketty, Thomas, and Emmanuel Saez. 2014. 'Inequality in the Long Run.' *Science* 344 (6186): 838–43.
36. Polanyi, Karl. 1944. *The Great Transformation*. New York, NY: Farrar & Rinehart, inc.
37. Rajan, Raghuram G, and Luigi Zingales. 2003. *Saving Capitalism from the Capitalists: Unleashing the Power of Financial Markets to Create Wealth and Spread Opportunity*. 1st ed. New York, NY: Crown Business.
38. Robison, Jennifer. 2011. 'Happiness Is Love - and $75,000.' *Gallup Business Journal*.
39. Seabright, Paul. 2010. *The Company of Strangers: A Natural History of Economic Life* (Revised Edition). Princeton, NJ: Princeton University Press.
40. Smith, Adam. (1776) 2003. *An Inquiry into the Nature and Causes of the Wealth of Nations*. New York, NY: Random House Publishing Group.
41. Smith, Adam. (1759) 2010. *The Theory of Moral Sentiments. Edited by Ryan Patrick Hanley*. New York, NY: Penguin Group.
42. Temin, Peter. 1997. 'Two Views of the British Industrial Revolution.' *The Journal of Economic History* 57 (01).
43. World Bank. 1993. *The East Asian Miracle: Economic Growth and Public Policy*. New York, NY: Oxford University Press.

2

TECHNOLOGICAL CHANGE, POPULATION AND ECONOMIC GROWTH



Shutterstock

HOW IMPROVEMENTS IN TECHNOLOGY HAPPEN, AND HOW THEY SUSTAIN GROWTH IN LIVING STANDARDS

- Economic models help explain the Industrial Revolution, and why it started in Britain
- Wages, the cost of machinery and other prices all matter when people make economic decisions
- In a capitalist economy innovation creates temporary rewards for the innovator, and this provides incentives for improvements in technology to reduce costs
- These rewards are destroyed by competition when innovation diffuses throughout the economy
- Population, the productivity of labour, and living standards may interact to produce a vicious circle of economic stagnation
- The permanent technological revolution associated with capitalism allowed some countries to make a transition to sustained growth in living standards

Institute for New Economic Thinking
UCL SciencesPo Azim Premji University
FRIENDS PROVIDENT Foundation

In 1845 a mysterious disease appeared for the first time in Ireland. It caused potatoes to rot in the ground; by the time it became clear that a plant was infected, it was too late. The potato blight, as it became known, devastated Irish food supplies for the rest of the decade. Starvation spread. By the time the Irish famine had ended about a million people had died, out of an initial total of 8.5 million, equivalent as a fraction of the population to the mortality suffered by Germany in defeat in the second world war.

The Irish famine sparked a worldwide relief effort. Former slaves in the Caribbean, convicts in Sing Sing prison in New York, Bengalis both rich and poor, and Choctaw Native Americans all donated money, as did such celebrities as the Ottoman Sultan Abdulmecid and Pope Pius IX. Then as now, ordinary people felt empathy for others who were suffering, and acted accordingly.

But many economists were much more hard-hearted. One of the best-known, Nassau Senior, consistently opposed British government famine relief, and was reported by a horrified Oxford University colleague as saying that "he feared the famine of 1848 in Ireland would not kill more than a million people, and that would scarcely be enough to do much good."

Senior's views are morally repulsive, but they did not reflect a genocidal desire to see Irish men and women die. Instead, they were a consequence of one of the most influential economic doctrines of the early 19th century, Malthusianism. This was a body of theory developed by an English clergyman, Thomas Robert Malthus, in *An Essay on the Principle of Population* first published in 1798.

Malthus held that a sustained increase in income per capita would be impossible. This was because even if technology improved, raising the productivity of labour, as soon as people were somewhat better off they would have more children. Then the growth in population would continue until living standards fell sufficiently to halt the increase in population. Malthus' vicious circle of poverty was widely accepted as inevitable. It provided an explanation of the world in which Malthus lived, in which incomes might fluctuate from year to year or even century to century, but not trend upwards. This had been the case in many countries for at least 700 years before Malthus published his essay, as we saw in Figure 1.1a.

Unlike Adam Smith, whose *Wealth of Nations* had appeared just 22 years earlier, Malthus' book did not offer an optimistic vision of economic progress—at least as far as ordinary farmers or workers were concerned. Even if people succeeded in improving technology, in the long run the vast majority of people would earn enough from their jobs or their farms to keep them alive, and no more.

But in Malthus' lifetime something big was happening all around him, changes that would soon allow Britain to escape from the vicious circle of population growth and stagnation of income that he described. The change that had sprung Britain from

the Malthusian trap, and would do the same for many countries in the 100 years that followed, is known as the *Industrial Revolution*—an extraordinary flowering of radical invention that allowed the same output to be produced with less labour.

In textiles the most famous inventions involved spinning (traditionally carried out by women known as spinsters, meaning female spinner, a term which has come to mean an older unmarried woman), and weaving (traditionally carried out by men). In 1733 John Kay invented the flying shuttle, which greatly increased the amount a weaver could produce in an hour. This increased the demand for the yarn that was used in weaving, to the point where it became difficult for spinsters to produce sufficient quantities using the spinning wheel technology of the day. James Hargreaves' spinning jenny, introduced in 1764, was a response to this problem.

Technological improvements in other areas were equally dramatic. James Watt's steam engine, introduced at the same time as Adam Smith published the *Wealth of Nations*, was typical. They were gradually improved over a long period of time. And they were eventually used across the economy: not just in mining, where the first steam engine powered water pumps, but also in textiles, manufacturing, railways and steamships. They are an example of what is termed a general-purpose innovation or technology. In recent decades the most obvious equivalent is the computer.

Coal played a central role and Great Britain had a lot of it. Prior to the Industrial Revolution, most of the energy used in the economy was ultimately produced by edible plants, which converted sunlight into food for both animals and people, or by trees whose wood could be burned or transformed into charcoal. By switching to coal, humans were able to exploit a vast reserve of what is effectively stored sunlight. The cost has been the environmental impact of burning fossil fuels, as we saw in Unit 1 and will return to in Unit 18.

These inventions, alongside other innovations of the Industrial Revolution, broke Malthus' vicious circle. Advances in technology raised the amount that a person could produce in a given amount of time, allowing incomes to rise even as population itself was increasing. And as long as technology continued improving quickly enough, it might outpace the growth in population that resulted from the increased income. Living standards could then rise. Much later, people would prefer smaller families, even when they earned enough to afford to have a lot of children.

This is what happened in Britain, and later in many parts of the world.

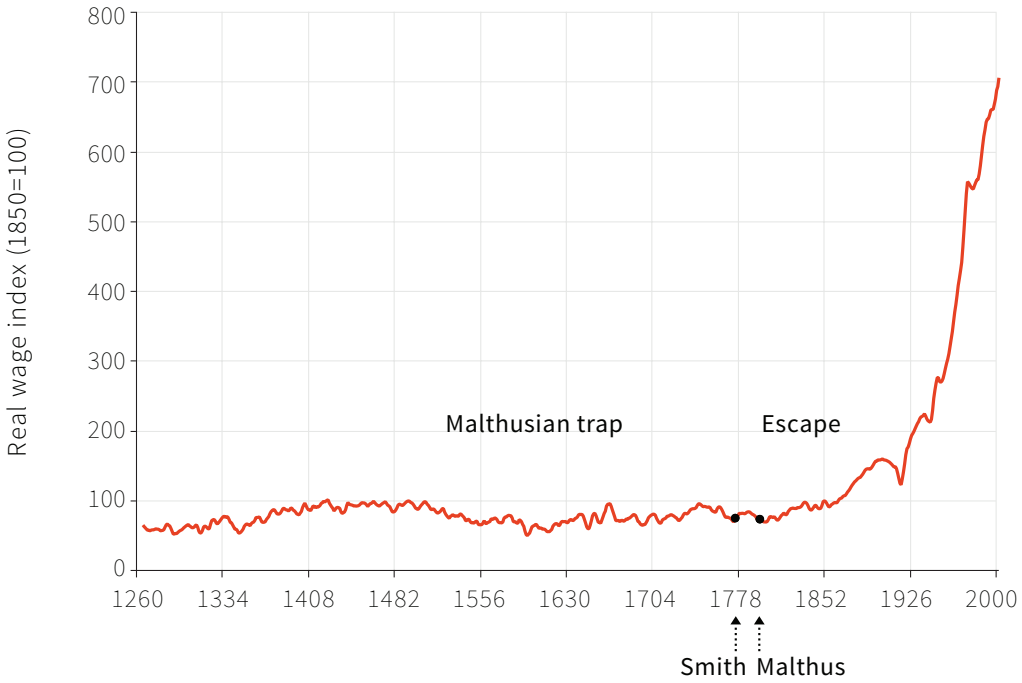


Figure 2.1 *Real wages over seven centuries: Craftsmen (skilled workers) in London (1264-2001)*

Source: *Methods used for calculating the data are covered in: Allen, Robert C. 2001. 'The Great Divergence in European Wages and Prices from the Middle Ages to the First World War.' Explorations in Economic History 38 (4): 411–47.*

Figure 2.1 shows an index of the average real wage of skilled craftsmen in London between the years 1264 and 2001. There is a long period in which living standards were trapped by Malthusian logic, and a dramatic increase that took place after 1830.

Why did the spinning jenny, the steam engine, and a cluster of other inventions emerge and spread across the economy in Britain at this time? This is one of the most important questions in economic history, and historians continue to argue about it.

In this unit we provide one explanation of how the improvements in technology came about, and why they occurred first only in Britain, and in the 18th century. We will also explain why the long flat part of Figure 2.1's hockey stick proved so hard to escape not only in Britain, but also throughout the world in the 200 years that followed. We will do this by building *models*: simplified representations that help us to understand what is going on by focusing attention on what is important. Models will help us understand both the kink in the hockey stick and the long flat handle.

2.1 ECONOMIC MODELS: HOW TO SEE MORE BY LOOKING AT LESS

What happens in the economy depends on what millions of people do, and the effects that their decisions have on the behaviour of others. It would be impossible to understand the economy by describing every detail of what they do and how they interact. We need to be able to stand back and look at the big picture. To do this we use models.

To create an effective model we need to distinguish between the essential features of the economy that are relevant to the question we want to answer, which should be built in to the model, and unimportant details that can be ignored.

Models come in many forms—and you have seen three of them already in Figures 1.8, 1.9 and 1.18 in Unit 1. For example, Figure 1.18 illustrated that economic interactions involve *flows* of goods (for example when you buy a washing machine), services (when you purchase haircuts or bus rides), and also people (when you spend a day working for an employer).

Figure 1.18 was a diagrammatic model illustrating the flows that occur within the economy, and between the economy and the biosphere. The model is not "realistic"—the economy and the biosphere don't look anything like it—but it nevertheless illustrates the relationships among them. The fact that the model omits many details—and in this sense is unrealistic—is a feature of the model, not a bug.

Some economists have used physical models to illustrate and explore how the economy works. For his 1891 PhD thesis at Yale University, Irving Fisher designed a hydraulic apparatus (Figure 2.2) to represent flows in the economy. It consisted of interlinked levers and floating cisterns of water to show how the prices of goods depend on the amount of each good supplied, the incomes of consumers, and how much they value each good. The whole apparatus stops moving when the

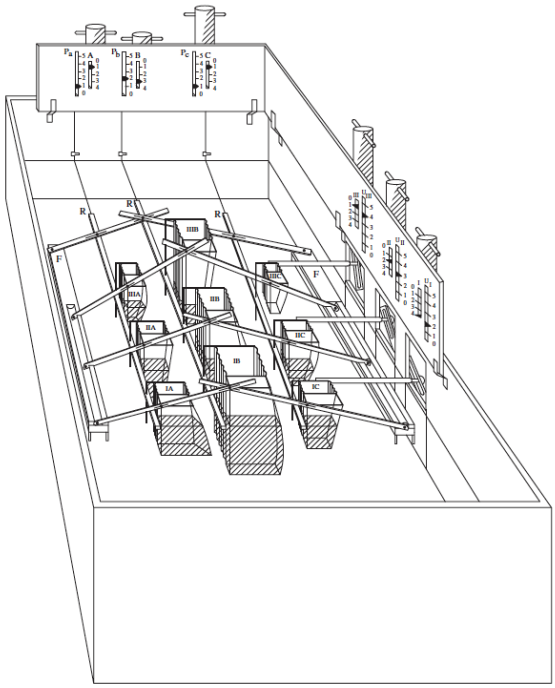


Figure 2.2 *Irving Fisher's sketch of his hydraulic model of economic equilibrium (1891).*

Source: *Brainard, William C., and Herbert E. Scarf. 2005. 'How to Compute Equilibrium Prices in 1891.' American Journal of Economics and Sociology 64 (1): 57–83*

water levels in the cisterns are the same as the level in the surrounding tank. When it comes to rest, the position of a partition in each cistern corresponds to the price of each good. For the next 25 years he would use the contraption to teach students how markets work.

How models are used in economics

Fisher's study of the economy illustrates how all models are used:

1. First he built a model to capture the elements of the economy that he thought mattered for the determination of prices.
2. Then he used the model to show how interactions between the elements could result in a set of prices that did not change.
3. Finally he conducted experiments with the model to discover the effects of changes in economic conditions: for example, if the supply of one of the goods increased, what would happen to its price? What would happen to the prices of all of the other goods?

Don't think that Irving Fisher was some kind of a crank, just because his doctoral dissertation represented the economy as a big tank of water. He went on to become one of the most highly regarded economists of the 20th century and his contributions formed the basis of modern theories of borrowing and lending that we will describe in Unit 11.

Fisher's machine illustrates an important concept in economics. An *equilibrium* is a situation that is self-perpetuating: that is, a situation in which something of interest does not change unless a force for change is introduced from the outside that alters the basic data describing the situation. Fisher's hydraulic apparatus represented equilibrium in his model economy by equalising water levels, which represented constant prices.

We will use the concept of equilibrium to explain prices in later units, but here we show that in the Malthusian model a wage equal to *subsistence* is an equilibrium. The reason is that, just like differences in the water levels in the different cisterns in Fisher's machine, movements away from subsistence wages are self-correcting: they automatically lead back to wages at a subsistence level.

Note: equilibrium means that one or more things are constant. It does not need to mean that nothing changes. We will also see that change (for example the rate of increase of GDP per capita, or the rate at which prices are increasing) can also be an equilibrium as long as it is self-perpetuating.

Although it is unlikely that you will build a hydraulic model for yourself, you will work with many existing models on paper or on a screen, and sometimes create your own models of the economy.

When we build a model, the process follows these steps:

1. We construct a *simplified description* of the conditions under which people take actions.
2. Then we describe in simple terms *what determines the actions* that people take.
3. We determine how each of their actions affects others.
4. We determine *the outcome of these actions*. This is often an equilibrium (something is constant).
5. Finally, we try to get more insight by studying *what happens when conditions change*.

Economic models often use mathematical equations and graphs as well as words. Mathematics is part of the language of economics, and it often helps in making our statements precise and easily understood by others. Much of the knowledge of economics, however, cannot be expressed simply by using mathematics. It requires clear descriptions, using standard definitions of terms.

We use mathematics as well as words to describe models. We will usually use graphs to do this; but, if you want, you will also be able to look at some of the equations behind the graphs. Just look for the references to our Leibniz features, and click on the links.

A model starts with some assumptions or hypotheses about how people behave, and often gives us some predictions about what we will observe in the economy. Gathering data on the economy, and comparing it with what a model predicts, helps us to decide whether the assumptions we made when we built the model—what to include, and what to leave out—were justified.

MODEL

What makes a good *model*?

- *It is clear*: It helps us better understand something important
- *It predicts accurately*: Its predictions are consistent with evidence
- *It improves communication*: It helps us to understand what we agree (and disagree) about
- *It is useful*: We can use it to find ways to improve how the economy works

Governments, central banks, corporations, trade unions and anyone else who makes policies or predicts the future use some type of simplified model.

Bad models often result in disastrous policies, as we will see later. To have confidence in those models, we need to confront them with evidence.

We will see that our economic models of the vicious cycle of Malthusian subsistence living standards and the permanent technological revolution pass this test—even though they leave many questions unanswered.

DISCUSS 2.1: DESIGNING A MODEL

For a country (or city) of your choice, look up a map of the railway or public transport network.

In designing this model, how do you think the modeller selected which features of reality to include?

2.2 BASIC CONCEPTS: PRICES, COSTS AND INNOVATION RENTS

We now introduce an economic model to help explain the circumstances under which new technologies are chosen, both in the past and in contemporary economies. The model also helps explain why some technologies that were replaced in the Industrial Revolution in Britain (such as the hand loom) are still in use today in some parts of the world.

We build the model using four key ideas of economic modelling:

- *Ceteris paribus* and other simplifications help us think clearly. We see more by looking at less.
- *Incentives* matter, because they affect the benefits and costs of taking one action as opposed to another.
- *Relative prices* help us compare alternatives.
- *Economic rent* is the basis of how we make choices.

Part of the process of learning to do economics is learning a new language. The terms below will recur frequently in the units that follow, and it is important to learn how to use them precisely and with confidence.

Ceteris paribus and simplification

As is common in scientific inquiry, economists often simplify our analysis by setting aside things that are thought to be of less importance, using the phrase "holding other things constant" or, more often, using the Latin expression *ceteris paribus* meaning "other things equal". For example, later in the course we simplify an analysis of what people would choose to buy by looking at the effect of changing a price—but ignoring influences on our behaviour like brand loyalty, or what others would think of our choices. These *ceteris paribus* assumptions, used well, can clarify the picture without distorting the key facts.

DISCUSS 2.2: WARNING: CETERIS PARIBUS CAN MISLEAD YOU!

A *ceteris paribus* assumption can be misleading, depending on the question being asked. In the example above of our shopping behaviour, think of questions for which the *ceteris paribus* assumptions (loyalty and self-image are not important in the model) might lead to a misleading conclusion.

When we study the way that a capitalist economic system promotes technological improvements, the kinds of things we might "hold constant" for the simplest possible model include differences from one town to the next in wages, similar differences in the prices of other inputs, different degrees of knowledge about technologies used in other firms, and the attitude of firm owners (or their managers) about taking risks. In other words, in this case we assume:

- Prices of all inputs are the same for all firms.
- All firms know the technologies in use in other firms.
- Attitudes towards risk are similar among firm owners.

Incentives matter

Why did the water in Fisher's hydraulic economy machine move when he changed the quantity of "supply" or "demand" for one or more of the goods, so that the prices were not longer in equilibrium?

- Gravity acts on the water so it finds the lowest level.
- Channels allow the water to seek out the lowest level, but restrict the ways in which it can flow.

All economic models have something equivalent to gravity and a description of the kinds of movements that are possible. The equivalent of gravity in economic models is the assumption that, when taking one course of action or another, people are attempting to do as well as they can (by some standard).

The analogy to the free movement of water in Fisher's machine is that people are free to select different courses of action, rather than simply being told the course they must take. This is where economic incentives affect the choices we make. But we can't do everything we want to do: not every channel is open to us.

Like many economic models, the one we use to explain the permanent technological revolution is based on the idea that people or firms respond to economic incentives. As we will see in Unit 4, people are motivated not only by the desire for material gain but also by love, hate, sense of duty, and desire for approval. But material comfort is definitely an important motive, and economic incentives appeal to this motive.

When owners or managers of firms decide how many workers to hire, or when shoppers decide what and how much to buy, prices are going to be an important factor determining their decision. If prices are a lot cheaper in the discount supermarket than in the corner shop, and it is not too far away, this will be a good argument for shopping in the supermarket rather than in the shop.

Relative prices

A third characteristic of many economic models is that we are interested in ratios of things, and not their absolute level. This is because economics focuses attention on alternatives and choices. For example, it is not the price level in one shop or the other that matters, but the price level in one shop compared to the other. In other words, relative prices matter for the decisions of shoppers, or consumers, as we tend to call them in economic theory. If supermarkets lowered their prices, but the corner shops then lowered their prices proportionally in response, there would be no incentive for consumers to switch away from convenience stores.

Relative prices are simply the price of one option relative to another. We often express relative price as the ratio of two prices. We will see that they matter a lot in explaining not just what consumers decide to buy, but why firms make the choices that they do.

Reservation positions and rents

To explain the great escape out of the flat part of the hockey stick, we need to understand the choices made by inventors and firms at the time of the Industrial Revolution. The prices that mattered most then were energy prices (the price of coal, for example, to power a steam engine), and the wage rate (the price of an hour of a worker's time). The ratio of the two—the price of coal relative to the price of labour—will play a big part in our story.

Imagine that you have figured out a new way of reproducing sound in high quality. Your invention is much cheaper to use than anyone else's method. Your competitors cannot copy you, either because they cannot figure out how to do it or because you have a patent on the process (making it illegal for them to copy you, even if they could). Suppose they continue offering their services at a price that is much higher than your costs.

If you match their price, or undercut them by just a bit, you will be able to sell as much as you can produce; so you just charge the same price, and make profits greatly in excess of what your competitors are making. In this case we say that you are making an *innovation rent*. Innovation rents are a form of economic rent—and economic rents occur throughout the economy. They are one of the reasons why capitalism can be such a dynamic system.

We will use the idea of innovation rents to explain some of the factors contributing to the Industrial Revolution. But *economic rent* is a general concept that will help explain many other features of the capitalist economy.

When taking some action (call it action A) results in your receiving a benefit greater than if you had chosen your next best action, we say that you have received an economic rent.

economic rent = benefit from option taken - benefit from next best option

The term is easily confused with such everyday uses of the word as in the rent for temporary use of a car, or apartment, or piece of land. To avoid this confusion, when we mean economic rent, we emphasise the word "economic". Remember, *an economic rent is something you would like to get, not something you have to pay*.

The alternative action (action B), with the next greatest net benefit, is often called the "next best alternative", your "reservation position" or, the term we use, *reservation option*. It is "in reserve" in case you do not choose *A*. Or, if you are enjoying *A* but then someone excludes you from doing it, your reservation option is your plan *B*, literally in this case. This is why you will also hear it called a "fallback option".

Economic rent gives us a simple decision rule:

- If action *A* would give you an economic rent (and if nobody else would suffer): *Do it!*
- If you are already doing action *A*, and it earns you an economic rent: *Carry on doing it!*

This decision rule lies behind our explanation of why the combination of private property, markets and firms promotes innovation. In the next section we will compare a particular technology *A* and technology *B*.

This can explain the escape from the Malthusian trap if we use the basic institutions of capitalism: private property, markets and firms. We show how when they are working well, as illustrated in Figure 1.11, these institutions provide the carrots and sticks stimulating both invention and diffusion of innovations.

- *Private property*: The fact that the firm is private property means that profits made after paying costs go to the owners, and will not be lost to either theft or government seizure. Firms that find a way to lower the cost of production without reducing quality stand to make substantial economic rents—until their competitors either copy them or find different ways to lower costs. These rents are the incentives driving the innovation process.
- *Markets*: The fact that the firm must compete in markets by selling goods at low cost means that those who fall behind will fail. This forces other firms either to copy the innovation of the first firm, or to lower costs in some other way. This provides a kind of economic version of Charles Darwin's natural selection that we can call the "survival of the profitable". If the rents made possible by innovation are the carrots, the prospect of failure for those who fall behind is the stick.
- *Firms*: The fact that most production is done in firms, rather than in families or governments, means that those who succeed can expand by attracting more funds to purchase capital goods, hiring more employees, and thereby benefiting their consumers and raising their owners' profits. Similarly, if the firm fails, it will eventually disappear.

The carrots and sticks driving the constant pressure to reduce costs are stronger when they operate through the combination of private property, markets, and firms than in economic systems lacking one or more of these institutions. The emergence of capitalist institutions in the 18th century thus created conditions for continuous technological progress, first in England, and then in other countries.

2.3 MODELLING A DYNAMIC ECONOMY: COST-REDUCING INNOVATION AND COMPETITION

Let us now apply these modelling ideas, and the description of how capitalism can promote innovation, to explain technological progress. We do this in three steps:

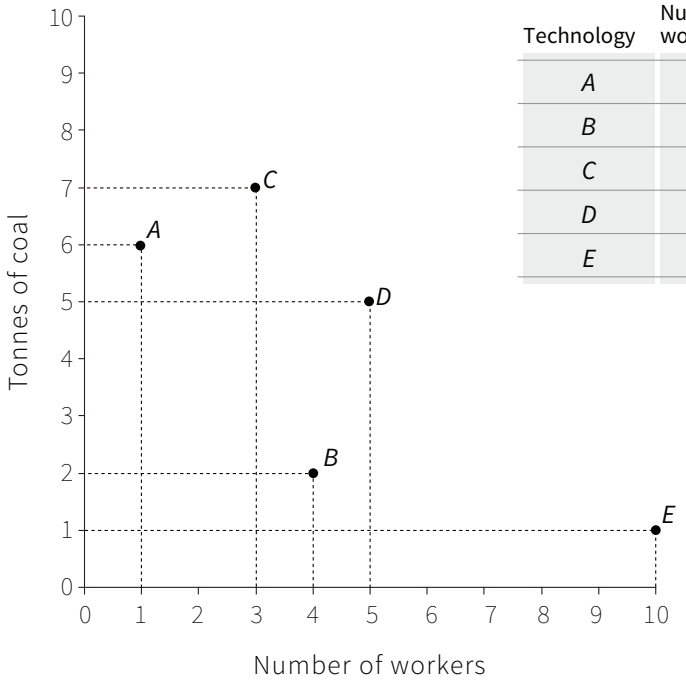
1. What is a technology?
2. How does a firm evaluate the cost of different technologies?
3. How does a successful innovation raise the profits of a firm?

What is a technology?

If we ask an engineer to report on the technologies that are available to produce 100 metres of cloth, where the inputs are labour (number of workers, each working for a standard work day, say 8 hours) and energy (tonnes of coal), the answer can be represented in the diagram and table in Figure 2.3. The five points in the table represent five different technologies. For example, technology *E* uses 10 workers and 1 tonne of coal to produce 100 metres of cloth.

We describe the *E*-technology as relatively labour-intensive and the *A*-technology as relatively energy-intensive. If an economy were using technology *E* and shifted to using technology *A* or *B* we would say that they had adopted a labour-saving technology, because the amount of labour used to produce 100 metres of cloth with these two technologies is less than with technology *E*. This is what happened during the Industrial Revolution.

Use the slideline in Figure 2.3 to represent the five technologies in the diagram.



Technology	Number of workers	Coal required (tonnes)
A	1	6
B	4	2
C	3	7
D	5	5
E	10	1

Figure 2.3 *Different technologies for producing 100 metres of cloth.*

Interact

Follow figures click-by-click in the full interactive version at www.core-econ.org.

Which technology will the firm choose? The first step is to rule out technologies that are obviously inferior. We begin in Figure 2.4 with the *A*-technology and look to see whether any of the alternative technologies use at least as much labour and coal. The *C*-technology is inferior to *A*: to produce 100 metres of cloth, it uses more workers (three rather than one) and more coal (7 tonnes rather than 6 tonnes). We say the *C*-technology is *dominated* by the *A*-technology: there are no conditions (no sets of positive prices, for example) that would ever induce a firm to use technology *C* when *A* is available. The slideline shows you a way to see which of the technologies are dominated, and which technologies dominate.

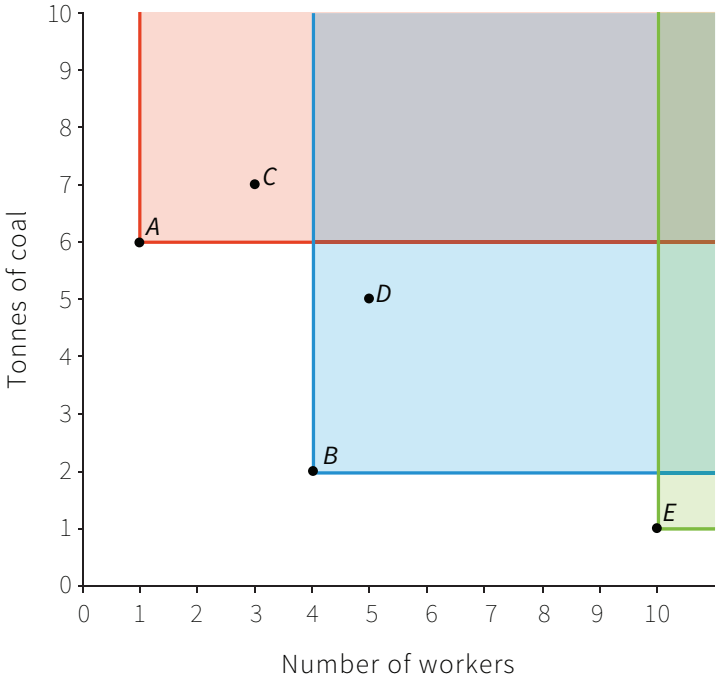


Figure 2.4 *Technology A dominates C; Technology B dominates D.*

Which technologies dominate others?

As in Figure 2.3, the five technologies for producing 100m of cloth are represented by the points *A* to *E*. We can use this figure to show which technologies dominate others. Clearly, technology *A* dominates the *C*-technology: the same amount of cloth can be produced using *A* with inputs of less labour and less energy. This means that, whenever *A* is available, you would never use *C*. Technology *B* dominates the *D*-technology: the same amount of cloth can be produced using *B* with inputs of less labour and less energy. Note *B* would dominate any other technology that is in the shaded area above and to the right of point *B*. The *E*-technology does not dominate any of the other available technologies. We know this because none of the other four technologies are in the area above and to the right of *E*.

Figure 2.4 has shown that the *C*- and *D*-technologies would never be chosen when *A* and *B* are available. Using only the engineering information about inputs, we have narrowed down the choice but how does the firm choose between *A*, *B* and *E*? This requires an assumption about what the firm is trying to do. We assume this goal is to make as much profit as possible, which means producing cloth at the least possible cost.

Making a decision about technology also requires economic information about relative prices—about the cost of hiring a worker, and of purchasing a tonne of coal. Intuitively, the labour-intensive *E*-technology would be chosen if labour was very cheap relative to the cost of coal; the energy-intensive *A*-technology would be preferable in a situation where coal is relatively cheap. An economic model helps us be more precise than this.

How does a firm evaluate the cost of production using different technologies?

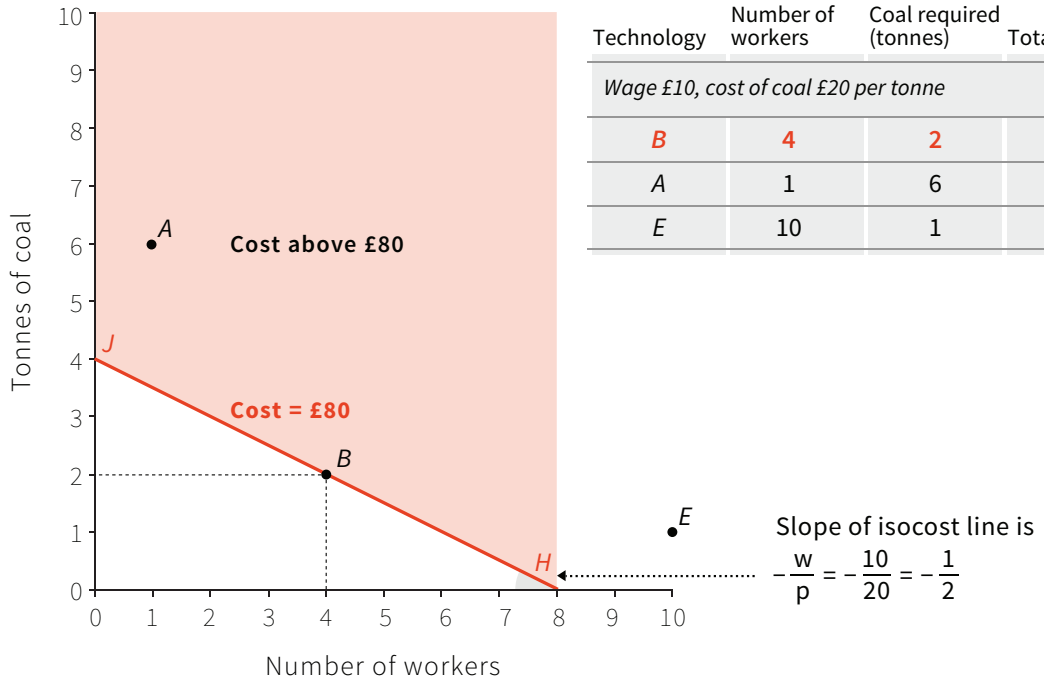
The cost of producing 100 metres of cloth using the three technologies that remain in play (that is, are not dominated) is calculated by multiplying the number of workers by the wage and the tonnes of coal by the price of coal. We use the symbol *w* for the wage, *L* for the number of workers, *p* for the price of coal and R for the tonnes of coal:

$$\begin{aligned} cost &= (wage \times workers) + (price\ of\ a\ tonne\ of\ coal \times number\ of\ tonnes) \\ &= (w \times L) + (p \times R) \end{aligned}$$

In the table in Figure 2.5, we have calculated the cost of producing 100 metres of cloth with technologies *A* and *B* when the wage is £10 and the price of coal is £20. Clearly that the *B*-technology allows the firm to produce cloth at lower cost. To show graphically that the *B*-technology produces cloth at lower cost than *A* and *E* when the wage is 10 and the price of coal is 20, we construct something called the *isocost* line. This is a line along which all the combinations of workers and coal cost the same amount, for example, £80. It is called isocost because *iso-* is the Greek for "same". Work through the steps in Figure 2.5 to see the cost of producing 100 metres of cloth with each technology, and draw the £80 isocost line.

To construct the isocost line for a total cost of £80, we ask how many workers could be employed for £80 (assuming zero outlay on coal). The answer is eight. This will be a point on the isocost line and is labelled *H*. If, instead, the £80 was used only on the purchase of coal, 4 tonnes could be bought. This gives point *J*. All the points on the line between *J* and *H* represent a total cost of £80. Notice that when drawing the isocost line, we simplify by assuming that fractions of workers and of coal, however small, can be purchased.

We can see from Figure 2.5 that the B-technology lies on the £80 isocost line. The other two available technologies lie above the £80 isocost line and will not be chosen if *B*-technology is available and if the relative prices are £10 and £20 for the wage and tonne of coal, respectively.



Technology	Number of workers	Coal required (tonnes)	Total cost (£)
Wage £10, cost of coal £20 per tonne			
B	4	2	80
A	1	6	130
E	10	1	120

Figure 2.5 *The cost of using different technologies to produce 100 metres of cloth: Low relative cost of labour.*

DISCUSS 2.3: ISOCOST LINES

1. Using the table in Figure 2.5, write down the equation of the isocost line that passes through points *A* and *E*.
2. Calculate the slope of this isocost line and compare it to the line passing through *B*.

We can now interpret the isocost line shown on the diagram as an equation. To do this, we write c for the cost of production: at whatever value it is fixed (for example, £80). We begin with the cost of production equation:

$$c = (w \times L) + (p \times R)$$
$$= wL + pR$$

To draw the isocost line, we want to express it in the form:

$$y = a + bx$$

Where a, which is a constant, is the intercept on the y-axis, and b is the slope of the line. In our model, tonnes of coal, R, are on the y-axis, the number of workers, L, is on the x-axis and we shall see that the slope of the line is the wage relative to the price of coal, $-w/p$. The isocost line slopes downward so the slope term in the equation ($-w/p$) is negative.

The isocost line is derived as follows:

$$c = wL + pR$$

Which can be written as:

$$pR = c - wL$$

And further rearranged as:

$$R = \frac{c}{p} - \frac{w}{p}L$$

This tells us that the isocost line for a cost of $c = 80$ has a vertical axis intercept of $80/20 = 4$ and a negative slope equal to the wage divided by the price of coal, $-w/p = -1/2$. This is the relative price of labour.

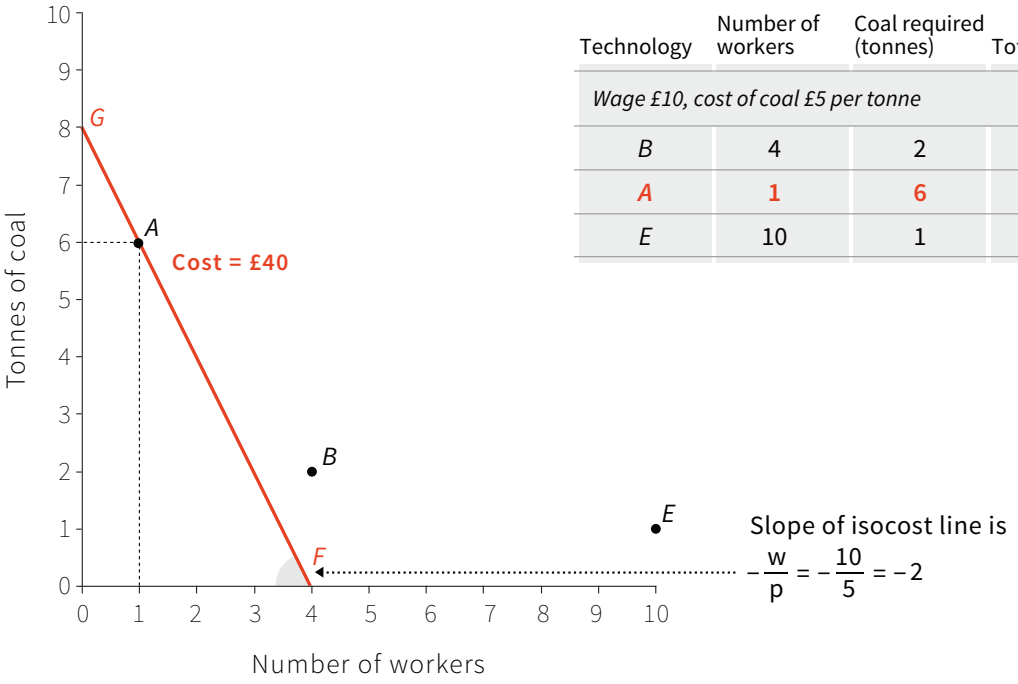
A wage is a special kind of price, so economists often refer to the wage as the price of (a given number of hours of) labour.

The effect of a change in relative prices

Any change in the relative price of the two inputs will change the slope of the isocost line. Looking at Figure 2.5, we can imagine that if the isocost line becomes sufficiently steep (with the wage rising relative to the cost of coal), the firm will switch to the A-technology. This is what happened in England in the 18th century.

Let's look at which technology will be least cost if the relative price of labour and coal change. Suppose that the price of coal falls to £5 and the wage remains at £10. Looking at the table in Figure 2.6, with the new prices, the *A*-technology allows the firm to produce 100 metres of cloth at least cost. Cheaper coal makes each method of production cheaper, but the energy-intensive technology is now cheapest.

To construct the isocost line going through point *A* with the new prices, we ask how many workers could be employed for £40 (assuming zero outlay on coal)—the answer is four. This will be a point on the isocost line and is labelled *F*. If, instead, the £40 was used only on the purchase of coal, 8 tonnes could be bought. This gives point *G*. The *A*-technology lies on the £40 isocost line. With these relative prices, the other two available technologies lie above the £40 isocost line and will not be chosen if the *A*-technology is available.



Technology	Number of workers	Coal required (tonnes)	Total cost (£)
Wage £10, cost of coal £5 per tonne			
B	4	2	50
A	1	6	40
E	10	1	105

Figure 2.6 *The cost of using different technologies for producing 100 metres of cloth: High relative cost of labour.*

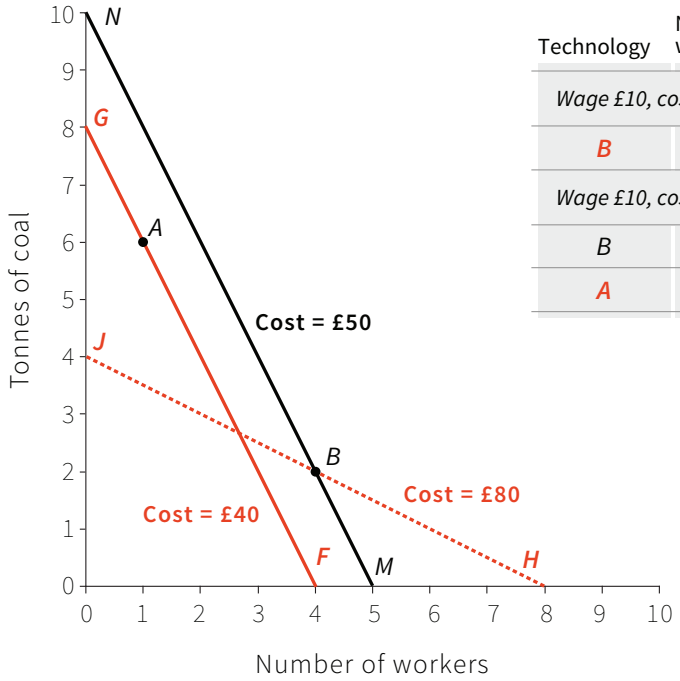
The slope of the isocost

The slope of the isocost line is equal to the wage divided by the cost of a tonne of coal. The slope is -2, because if you spent £20 less on coal by buying 4 tonnes less, and increased spending on labour by hiring two workers, total cost would remain unchanged at £40.

How does a cost-reducing innovation raise the profits of the firm?

The next step is to calculate the gains to the first firm to identify the least cost technology (*A*) when the relative price of labour to coal rises. Like all its competitors, the firm is initially using the *B*-technology and minimising its costs: this is shown in Figure 2.7 by the dashed isocost line through *B* (with end points *H* and *J*).

Once the relative prices change, the new isocost line through the *B*-technology is steeper and the cost of production is £50. A switch to the *A*-technology, which uses energy more intensively and labour less intensively to produce 100 metres of cloth, reduces costs to £40. Use the slideline to see how isocost lines change with the new relative prices.



Technology	Number of workers	Coal required (tonnes)	Total cost (£)
Wage £10, cost of coal £20 per tonne			
B	4	2	80
Wage £10, cost of coal £5 per tonne			
B	4	2	50
A	1	6	40

Figure 2.7 *The cost of using different technologies for producing 100 metres of cloth.*

The firm's profits are equal to the revenue it gets from sales of output minus its costs. Whether the new or old technology is used, the same price has to be paid for labour and coal and the same price is received for selling 100 metres of cloth.

The change in profit is equal to the fall in costs associated with adopting the new technology: profits rise by £10 per 100 metres of cloth:

$$\begin{aligned} \textit{profit} &= \textit{revenue} - \textit{costs} \\ \textit{change in profit, using A rather than B} &= \textit{change in revenue} - \textit{change in costs} \\ &= 0 - (40 - 50) \\ &= 10 \end{aligned}$$

Therefore, in this case, *economic rent* = 10.

Although cheaper coal made the B-technology cheaper than before, there is still an incentive to switch to technology A and make additional profits of £10 per 100 metres of cloth.

Your economic rent, if you adopt the new technology, is exactly the cost reduction made possible by the new technology. The decision rule—*if the economic rent is positive, do it!*—tells you to innovate.

In our example, although the engineer could describe the A-technology, it was not in use until a first adopter firm responded to the incentive created by the increase in the relative price of labour. The first adopter is called an *entrepreneur*. When we describe a person or firm as entrepreneurial, it refers to a willingness to try out new technologies and to start new businesses.

The economist Joseph Schumpeter (see below) made the adoption of technological improvements by entrepreneurs a key part of his explanation for the dynamism of capitalism. This is why innovation rents are often called *Schumpeterian rents*.

Innovation rents will not last forever. Other firms, noticing that entrepreneurs are making economic rents, introduce the new technology. They will also reduce their costs and their profits will increase.

In this case, with higher profits per 100 metres of cloth, the lower-cost firms will thrive. They will increase their output of cloth. As more firms introduce the new technology, the supply of cloth to the market increases and the price will start to fall. This process will continue until everyone is using the new technology, at which stage prices will have declined to the point where no one is earning innovation rents. The firms that stuck to the old *B*-technology will be unable to cover their costs at the new lower price for cloth, and they will go bankrupt. Joseph Schumpeter called this *creative destruction*.

GREAT ECONOMISTS

JOSEPH SCHUMPETER

Joseph Schumpeter (1883-1950) developed one of the most important concepts of modern economics: *creative destruction*.

Schumpeter brought to economics the idea of the entrepreneur as the central actor in the capitalist economic system. The entrepreneur is the agent of change who introduces new products, methods of production and opens up new markets. Imitators follow and the innovation is diffused through the economy. A new entrepreneur and innovation launch the next upswing.

For Schumpeter, creative destruction was the essential fact about capitalism: old technologies and the firms that do not adapt are swept away by the new, because they cannot compete in the market by selling goods at a price that covers the cost of production. The failure of unprofitable firms releases labour and capital goods for use in new combinations.

This decentralised process generates a continued improvement in productivity, which leads to growth, so Schumpeter argued it is virtuous. Both the destruction of old firms, and the creation of new ones, takes time. The slowness of the process creates upswings and downswings in the economy. The branch of economic thought known as *evolutionary economics* (you can read articles on the subject here) can clearly trace its origins to Schumpeter's work, as well as most modern economic modelling that deals with entrepreneurship and innovation. Read Schumpeter's ideas and opinions in his own words, and an online essay about his work by Robert Skidelsky, a historian of economic thought.

Schumpeter was born in Austro-Hungary, but migrated to the US after the Nazis won the election in 1932 that led to the formation of the Third Reich in 1933. He had also experienced the first world war and the depression of the 1930s, and died while writing an essay called *The march into socialism*, recording his concerns about the increasing role of government in the economy and the resulting "migration of people's economic affairs from the private into the public sphere."

As a young professor in Austria he had fought and won a duel with the university librarian to ensure that students had access to books. He also claimed that as a young man he had three ambitions in life: to become the world's greatest economist, the world's greatest lover, and the world's greatest horseman. He added that only the decline of the cavalry had stopped him from succeeding in all three.

2.4 THE BRITISH INDUSTRIAL REVOLUTION AND THE INCENTIVE TO INTRODUCE NEW TECHNOLOGIES

What did inventions such as the spinning jenny do? The first spinning jennies had eight spindles. One machine operated by just one adult therefore replaced eight spinsters, working on eight spinning wheels. By the late 19th century a single spinning mule, operated by a very small number of people, could replace more than 1,000 spinsters. These machines did not rely on human energy, but were powered by water wheels or later coal-powered steam engines.

The old technology used a lot of workers, each with a small amount of machinery. The new technology used a lot of capital goods such as spinning mules, factory buildings, water wheels or steam engines. Like technology A in the example, they used a lot of coal and not much labour.

Before the Industrial Revolution technology was *labour-intensive*, capital goods-saving and energy-saving, while the new technology was *capital goods-intensive*, energy-intensive and labour-saving.

The model in the previous section provides a *hypothesis* (potential explanation) about why someone would bother to invent such a technology, and why someone would want to use it. In the model, we used two-dimensional graphs, which showed the producers of cloth choosing between technologies using just two inputs—energy and labour. This is a simplification, but it highlights the role of changes in the relative costs of inputs when we choose a technology. When the cost of energy increased relative to the cost of labour, this increased the cost of producing cloth using the old technology. It meant that there were innovation rents to be earned from a switch to the energy-intensive technology.

This is just one hypothesis. Is it actually what happened? Looking at how relative prices differed among countries, and how they changed over time, can help us understand why labour-saving, capital-using and energy-using technologies of the Industrial Revolution were invented in Britain rather than elsewhere. We can use the same reasoning to explain why it did not happen many centuries before.

No doubt it helped that Britain was such an inventive country. There were many skilled workmen, engineers and machine makers who could build the machines that inventors designed.

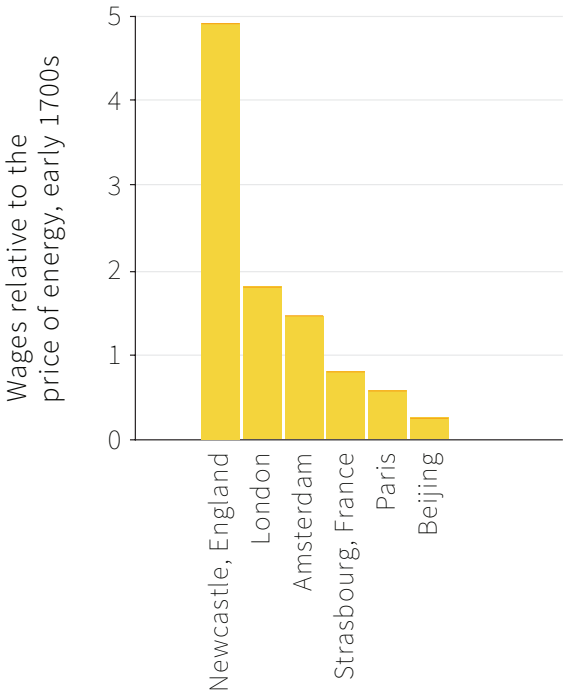


Figure 2.8 *Wages relative to the price of energy, (early 1700s).*

Source: *Page 140 of Allen, Robert C. 2008. The British Industrial Revolution in Global Perspective. Cambridge: Cambridge University Press.*

Figure 2.8 shows the price of labour relative to the price of energy in various cities in the early 1700s. In other words, it shows the wages of building labourers divided by the price of energy (more precisely, the price of 1 million British Thermal Units (BTU), a unit of energy equivalent to slightly more than 1,000 *joules*). You can see that labour was very expensive relative to the cost of energy in England and the Netherlands. It was less expensive in France (Paris and Strasbourg), and a lot less expensive in China.

Wages were high in England, relative to the cost of energy, both because English wages were higher than wages elsewhere, and because coal was cheaper in coal-rich Britain than in the other countries in the chart.

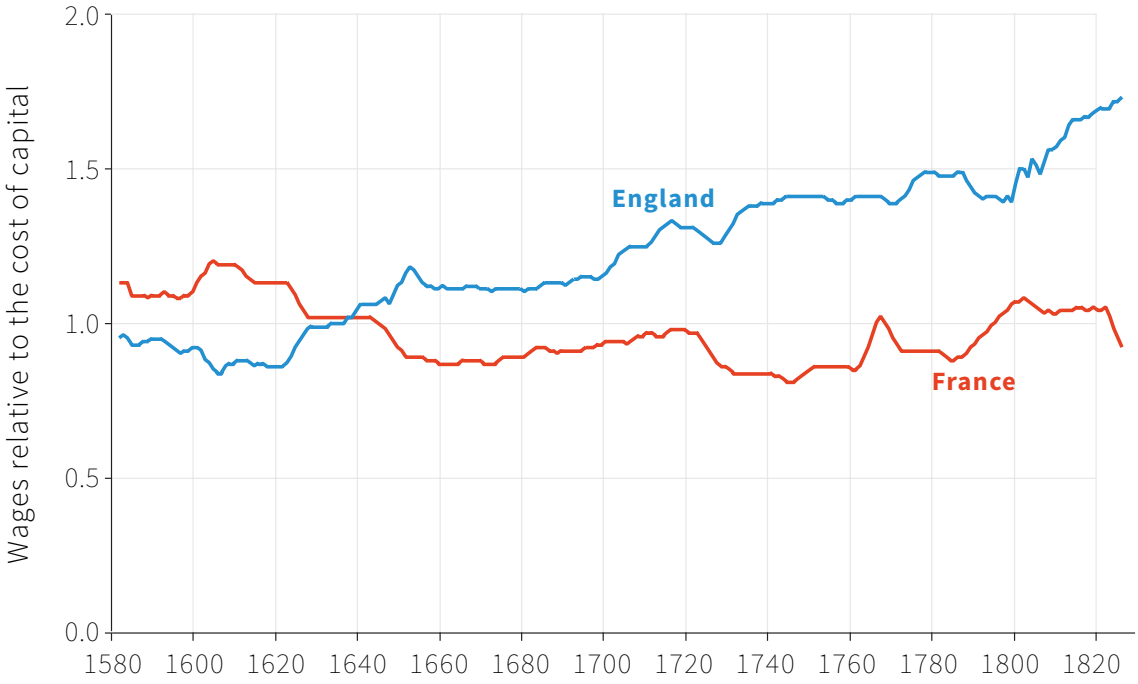


Figure 2.9 *Wages relative to the cost of capital goods (late 16th to the early 19th century).*

Source: *Page 138 in Allen, Robert C. 2008. The British Industrial Revolution in Global Perspective. Cambridge: Cambridge University Press.*

Figure 2.9 shows trends in the cost of labour relative to the cost of capital goods, in England and France from the late 16th to the early 19th century. It shows the wages of building labourers divided by the cost of using capital goods. This cost is calculated from the prices of metal, wood and brick, and the cost of borrowing, and takes account of the rate at which the capital goods wear out, or depreciate.

As you can see, wages relative to the cost of capital goods in England and France were similar in the mid 17th century but from then on, in England but not in France, workers became steadily more expensive relative to capital goods. In other words, the incentive to replace workers with machines was increasing in England during this time, but the same was not true in France. In France the incentive to save labour by innovating had been stronger during the late 16th century than it was 200 years later, at the time the Industrial Revolution began to transform Britain.

From the model in the previous section we learned that the technology chosen depends on relative prices of the inputs. By combining the predictions of the model with the historical data, we have one explanation for the timing and location of the Industrial Revolution.

- Wages relative to the cost of energy and capital goods *went up in the 18th century in Britain*. This had not been the case in earlier historical periods.
- Wages relative to the cost of energy and capital goods *were higher in Britain during the 18th century than elsewhere*.

Figure 2.10 uses the model to illustrate the British case.

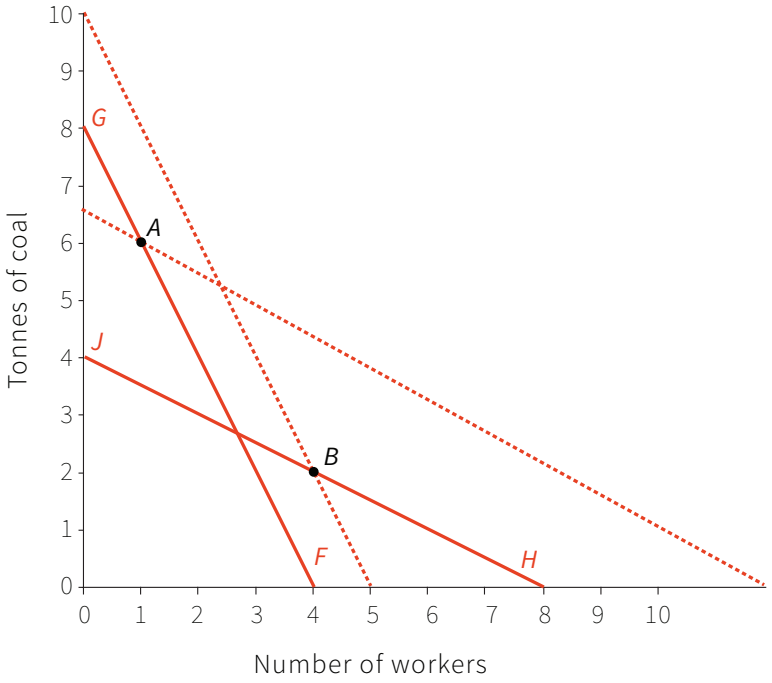


Figure 2.10 *The cost of using different technologies for producing 100 metres of cloth in Britain in the 17th and 18th centuries.*

Technology in the 1600s

In the 1600s, the relative prices are shown by isocost line HJ. The B-technology was used. At those relative prices, there was no incentive to develop a technology like A, which is outside the isocost line HJ. In the 1700s, the isocost lines such as FG were much steeper, because the relative price of labour to coal was higher. The relative cost was sufficiently high to make the A-technology lower cost than the B-technology. We know that, when the relative price of labour is high, technology A is lower cost because the B-technology lies outside the isocost line FG.

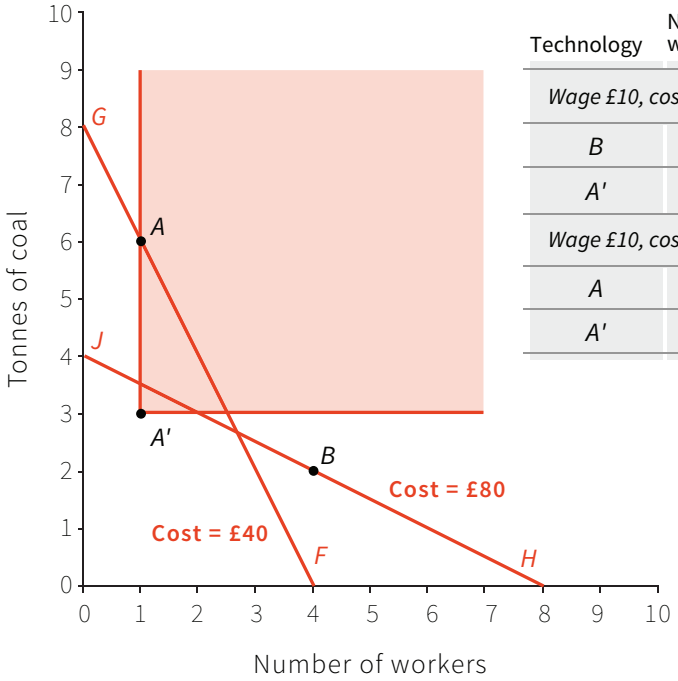
DISCUSS 2.4: BRITAIN BUT NOT FRANCE

Watch this video, in which Bob Allen, an economic historian, explains his theory of why the Industrial Revolution occurred when and where it did.

1. Summarise Allen's claim using the concept of economic rents. Which *ceteris paribus* assumptions are you making?
2. What other important factors may explain the rise of energy-intensive technologies in Britain in the 18th century?

England was a high-wage, cheap-energy and cheap-capital goods country, and so it makes sense that the energy-using, capital goods-using, labour-saving technologies of the Industrial Revolution were first adopted there. As a consequence, during the early years of the Industrial Revolution, technology advanced more rapidly in England than on the continent of Europe, and even more rapidly than in Asia.

What explains the eventual adoption of these new technologies in countries like France and Germany, and ultimately China and India? One answer is further technological progress, where a new technology is developed that dominates the existing one in use. Technological progress would mean that it would take smaller quantities of inputs to produce 100 metres of cloth. We can use the model to illustrate this. In Figure 2.11, technological progress leads to the invention of a superior energy-intensive technology, which is labelled *A'*. The slideline shows that the *A'*-technology dominates the A-technology. Once *A'* is available, the *A*-technology would no longer be chosen in countries using *A*.



Technology	Number of workers	Coal required (tonnes)	Total cost (£)
Wage £10, cost of coal £20 per tonne			
B	4	2	80
A'	1	3	70
Wage £10, cost of coal £5 per tonne			
A	1	6	40
A'	1	3	25

Figure 2.11 *The cost of using different technologies for producing 100 metres of cloth.*

Notice also that the new technology makes it profitable to switch even in a low-wage, expensive-energy economy. In spite of a low relative price of labour, the*A'*-technology is lower cost than the labour-intensive *B*-technology: *A'* is inside the *HJ* isocost line.

A second factor that promoted the diffusion across the world of the new technologies was wage growth and falling energy costs (due for example to cheaper transportation, allowing countries to import energy cheaply from abroad). This made isocost lines steeper in poor countries. And, as we have seen, if the isocost line becomes steep enough, energy-intensive technology is the least-cost method of production.

Either way, the new technologies started to spread, and the divergence in technologies and living standards was eventually replaced by convergence—at least among those countries in which the capitalist revolution had taken off.

Nevertheless, in some countries, we can still observe the use of technologies that were replaced in Britain in the Industrial Revolution. The model predicts that the relative price of labour must be very low in such situations, making the isocost line very flat. The *B*-technology could still be preferred in Figure 2.11 even when the *A'*-technology is available if the isocost line is even flatter than *HJ*, so that it goes through *B* but is below *A'*.

2.5 MALTHUSIAN ECONOMICS: DIMINISHING AVERAGE PRODUCT OF LABOUR

The historical evidence supports our model of how relative prices and innovation rents can provide a simple account of both the timing and the geographical spread of the permanent technological revolution.

This is part of the explanation of the upward kink in the hockey stick. Explaining the long flat part of the stick is another story, which requires a different model.

Malthus provides a model of the economy that predicts a pattern of economic development consistent with the flat part of the GDP per capita hockey stick from Figure 1.1a in Unit 1. His model introduces concepts that are used to analyse many other economic problems. One of the most important concepts in economics is the idea of diminishing average product of a factor of production.

Diminishing average product of labour

To understand what this means, imagine an agricultural economy that produces just one good, which we will call "food". Imagine that food production is very simple—it involves only farm labour, working on the land. In other words, ignore the fact that food production also requires spades, combine harvesters, chicken coops, grain elevators, silos, and every other type of building and equipment.

Labour and land (and the other inputs that we have ignored for now) are called factors of production, meaning inputs into the process of production. In the model of technological change above, the factors of production are energy and labour.

We will use one further *ceteris paribus* assumption to simplify. We assume that the amount of land is fixed, all of the same quality. Imagine that the land is divided into 800 farms, each worked by a single farmer. Each farmer works the same total hours during a year. Together these 800 farmers produce a total of 500,000kg of grain. The *average product* of a farmer's labour is:

$$\text{average product of labour} = \frac{\text{total output}}{\text{total number of farmers}}$$
$$= \frac{500{,}000\text{kg}}{800 \text{ farmers}}$$
$$= 625\text{kg per farmer}$$

To understand what will happen when population grows so that there are more farmers on the same limited space of farmland, we need something that economists call the *production function* for farming. This indicates the amount of output that will result for each number of farmers working the given amount of land. In this case we are holding constant all of the other inputs, including land, so we only consider how output varies with the amount of labour.

PRODUCTION FUNCTION

The amount of output that will result for one or more combinations of input. A *production function* describes differing technologies capable of producing the same thing.

In the previous sections you have already seen very simple production functions that specified the amount of both labour and energy necessary to produce 100 metres of cloth. For example, in Figure 2.3, the "cloth production function" says that using technology *B*, if 4 workers and 2 tonnes of coal are put into production, 100 meters of cloth will be the output. For technology *A*, the production function gives us another "if-then" statement: if 1 worker and 6 tonnes of coal are put into production then 100 meters of cloth will be the output.

The average product of labour in technology *B* is 25 metres of cloth and in technology *A* is 100 metres of cloth per worker. Economists conventionally do not include in the production function technologies that would never be used (no matter what were the relative prices of inputs). So technologies *C* and *D* in Figure 2.3 are not part of the "cloth production function" because while they are feasible from an engineering standpoint, they are irrelevant economically.

The grain production function is a similar "if-then" statement indicating that if there are *X* farmers then they will harvest *Y* grain.

Figure 2.12a lists some values of labour input and grain production. In Figure 2.12b, we assume that the relationship holds for all the farmers and grain production figures in between those shown in the table to draw the complete function.

We call this a production *function* because a function is a relationship between two quantities (inputs and outputs in this case), expressed mathematically as:

$Y = f(X)$

We say "Y is a function of X". X in this case is the amount of labour devoted to farming. Y is the output in grain that results from this input. The function $f(X)$ describes the relationship between the two, represented by the curve in the figure. Read our Leibniz supplement if you would like an introduction to the mathematical representation of production functions. The Leibniz supplements will introduce you to more sophisticated mathematics and calculus. You don't need them to understand our models, but they may help you if you are taking more advanced courses or studying mathematics as part of your economics course.

LABOUR INPUT (Number of workers)	GRAIN OUTPUT (kg)	AVERAGE PRODUCT OF LABOUR (kg/worker)
200	200,000	1,000
400	330,000	825
600	420,000	700
800	500,000	625
1,000	570,000	570
1,200	630,000	525
1,400	684,000	490
1,600	732,000	458
1,800	774,000	430
2,000	810,000	405
2,200	840,000	382
2,400	864,000	360
2,600	882,000	340
2,800	894,000	319
3,000	900,000	300

Figure 2.12a *Recorded values of a farmer's production function: Diminishing average product of labour.*

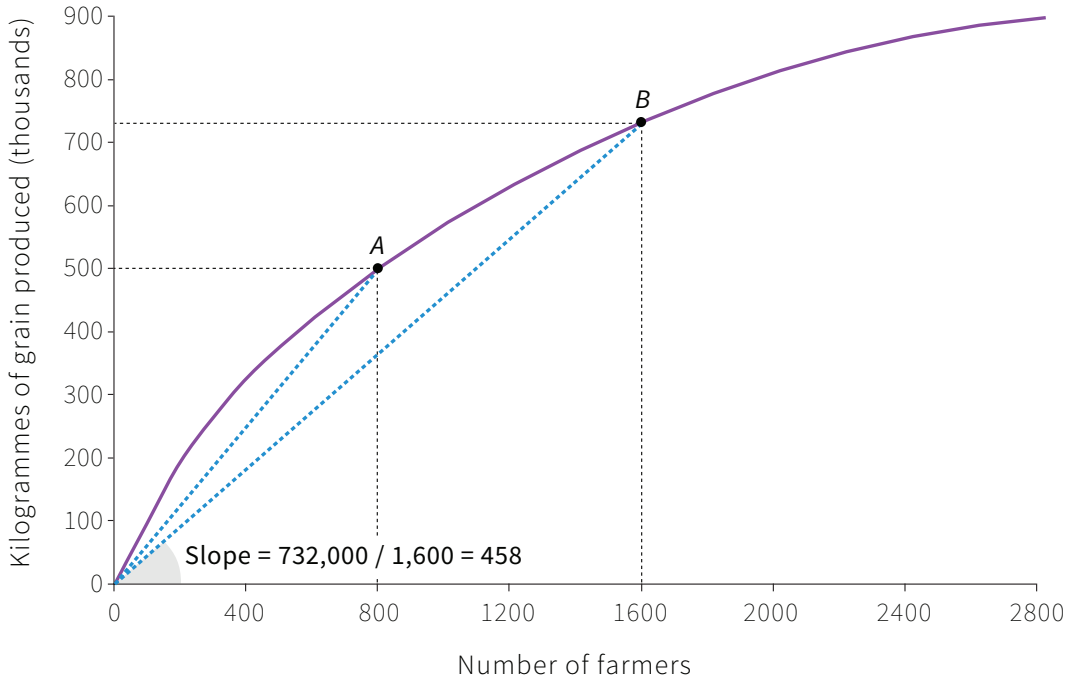


Figure 2.12b *The farmer's production function: Diminishing average product of labour.*

The farmer's production function

The farmer's production function: how the amount of farmers working the land translates into grain produced at the end of the growing season. Point A on the production function shows the output of grain produced by 800 farmers. Point B on the production function shows the output of grain produced by 1,600 farmers. At A the average product of labour is 500,000/800 = 625kg of grain per farmer. At B the average product of labour is 732,000/1,600 = 458kg of grain per farmer. The slope of the ray from the origin to point B on the production function shows the average product of labour. The slope is 458, meaning an average product of 458kg per farmer when 1,600 farmers work the land. The slope of the ray to point A is steeper than to B. When only 800 farmers work the land there is a higher average product of labour. The slope is 625, the average product of 625kg per farmer that we calculated previously.

DISCUSS 2.5: THE FARMER'S PRODUCTION FUNCTION

In Unit 1 we explained that the economy is part of the biosphere. Think of farming biologically. The function shown is hypothetical, but realistic if the farmer has a small plot of land.

1. Find out how many calories a farmer burns, and how many calories are contained in 1kg of grain.
2. Does farming produce a surplus of calories—more calories in the output than used up in the work input—using the production function in Figure 2.12b?

Note two things about the grain production function:

- *Labour combined with land is productive*. No surprises there. The more farmers there are, the more grain is produced; at least up to a point (3,000 farmers, in this case) beyond which putting an extra farmer on to the land would not produce more grain.
- *As more farmers work the land, the average product of labour falls*. The *diminishing average product of labour* is one of the two foundations of Malthus' model.

The average product of labour is the grain output divided by the amount of labour input. From the production function in Figure 2.12b, or from the table in Figure 2.12a (it's the same information) we see that an annual input of 800 farmers working the

land brings an average per farmer output of grain of 625kg, while increasing the labour input to 1,600 farmers produces an average per hour output per farmer of 458kg. The average product of labour falls as more labour is expended on production.

This worried Malthus.

To see why he was worried, imagine that, a generation later, each farmer has had many children, so that instead of a single farmer working each farm, there were now two farmers working. The total labour input into farming was 800, but is now 1,600. Instead of a harvest of 625kg of grain per farmer, the average harvest is now only 458kg.

You might argue that, in the real world, as the population grows more land can be used for farming. But Malthus pointed out that the first generation of farmers would have picked the best land, so any new land would be worse. This also reduces the average productivity of labour.

This means that diminishing average product of labour can be caused by:

- More labour devoted to a fixed quantity of land
- More (inferior) land brought into cultivation

Because the average product of labour diminishes as more labour is devoted to farming, as long as people do work more hours, their incomes inevitably fall.

2.6 MALTHUSIAN ECONOMICS: POPULATION GROWS WHEN LIVING STANDARDS INCREASE

On its own, the diminishing average product of labour does not explain the long, flat portion of the hockey stick. All the concept says is that living standards depend on the level of population. It doesn't say anything about why, over long periods, living standards and population didn't change much. For this we need the other part of Malthus's model: his argument that increased living standards create an increase in the population.

Malthus was not the first person to have this idea. Years before Malthus developed his theories, Richard Cantillon, an Irish economist, had stated that "Men multiply like mice in a barn if they have unlimited means of subsistence."

Malthusian theory essentially regarded people as being not that different from other animals:

"Elevated as man is above all other animals by his intellectual facilities, it is not to be supposed that the physical laws to which he is subjected should be essentially different from those which are observed to prevail in other parts of the animated nature."
—Thomas Robert Malthus, *A Summary View on the Principle of Population* (1830)

So the two ideas central to Malthus' model are:

- The law of diminishing average product of labour
- Population expands if living standards increase

Imagine a herd of antelopes on a vast and otherwise empty plain. Imagine also that there are no predators to complicate their lives (or our analysis). When these antelopes are better fed, they live longer and have more offspring. When the herd is small the antelopes can eat all they want, and the herd gets larger.

Eventually the herd will get so large relative to the size of the plain that the antelopes can no longer eat all they want. As the amount of land per animal declines, their living standards will start to fall. This reduction in living standards will continue as long as the herd continues to increase in size.

Since each animal has less food to eat, the antelopes will have fewer offspring and die younger; population growth will slow down. Eventually, living standards will fall to the point where the herd is no longer increasing in size. The antelopes have filled up the plain. At this point, each animal will be eating an amount of food that we will define as the *subsistence level*. When the animals' living standards have been forced down to subsistence level as a result of population growth, the herd is no longer getting bigger.

If antelopes eat less than the subsistence level, the herd starts to get smaller. And, as we have already seen, when consumption exceeds the subsistence level, the herd grows.

Much the same logic would apply, Malthus reasoned, if we considered a human population living in a country with a fixed supply of agricultural land. As long as people have "unlimited subsistence" they would multiply like Cantillon's mice in a barn; but eventually they would fill the country, and further population growth would push down the incomes of most people as a result of diminishing average product of labour. Falling living standards would slow population growth, as death rates increased and birth rates fell; ultimately incomes would settle at the subsistence level.

Malthus's model results in an *equilibrium* in which there is an income level just sufficient to allow a subsistence level of consumption. The variables that are unchanging in this equilibrium are:

- The population
- The income level of the people who make up the population

DISCUSS 2.6: ARE PEOPLE REALLY LIKE OTHER ANIMALS?

Malthus wrote: "[I]t is not to be supposed that the physical laws to which [mankind] is subjected should be essentially different from those which are observed to prevail in other parts of the animated nature."

Do you agree? Explain your reasoning.

Malthusian economics: The effect of a good harvest

In Figure 2.13 we illustrate how the combination of diminishing average product of labour, and the effect of higher incomes on population growth, mean that in the very long run, the income of farmers will not rise. In the figure, things on the left are causes of things to the right.

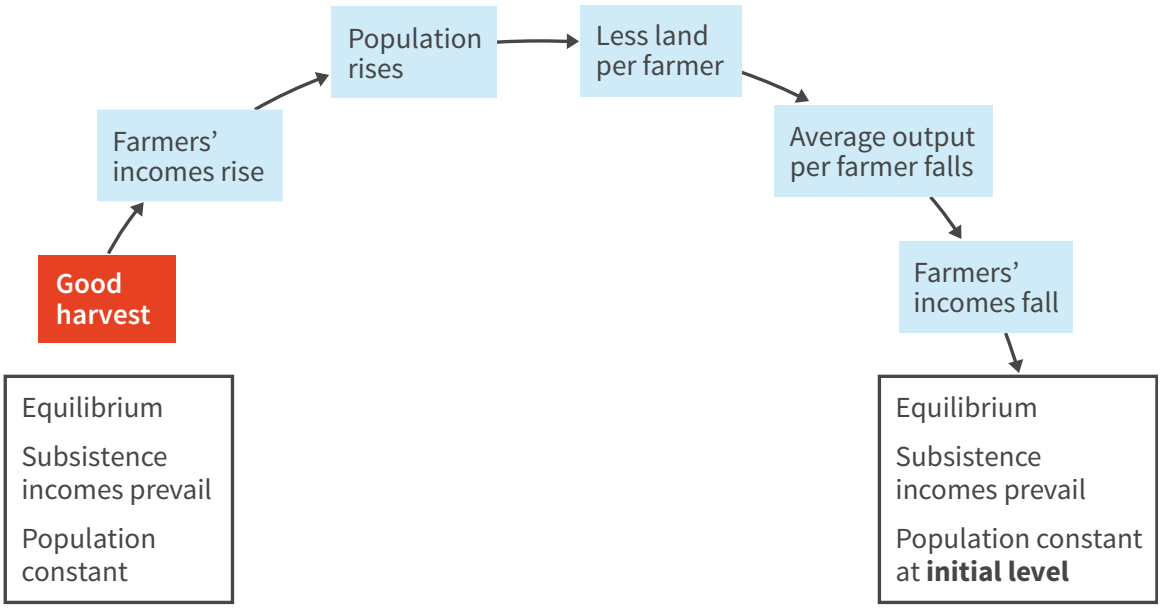


Figure 2.13 *Malthus' model: The effect of a good harvest.*

On the left side of the figure we have an economy with a constant population and with incomes at the subsistence level. This equilibrium state is then disrupted by a good harvest. Malthus reasoned that this raises farmers' incomes and so leads to a higher population. The next year we have:

- More people on the land
- A normal harvest

The result? Lower *average output* per farmer. This reduces the income of farmers, resulting in a decline in population in subsequent years, with the reduction in income continuing until it reaches subsistence levels and the population (by the definition of subsistence) is unchanging. It will also be at its original level, because that was the size of population that led to an average product of labour that matched subsistence incomes. In Figure 2.13 we reach the box on the right, which is exactly the equilibrium with which we began.

Malthusian economics: The effect of technological progress

Our next step is to use Malthus' model to predict the consequences of technological progress. We know that over the centuries before the Industrial Revolution, improvements in technology occurred in many regions of the world, including Britain, and yet living standards remained constant. The model predicts that there would have been a self-correcting response to new technology.

This is Figure 2.14. Beginning from equilibrium, with income at the subsistence level, a new technology such as an improved seed raises income per person on the existing fixed quantity of land. Higher living standards lead to an increase in population. As more people are added to the land, diminishing average product of labour means average income per person falls. Eventually incomes return to subsistence level, with a higher population.

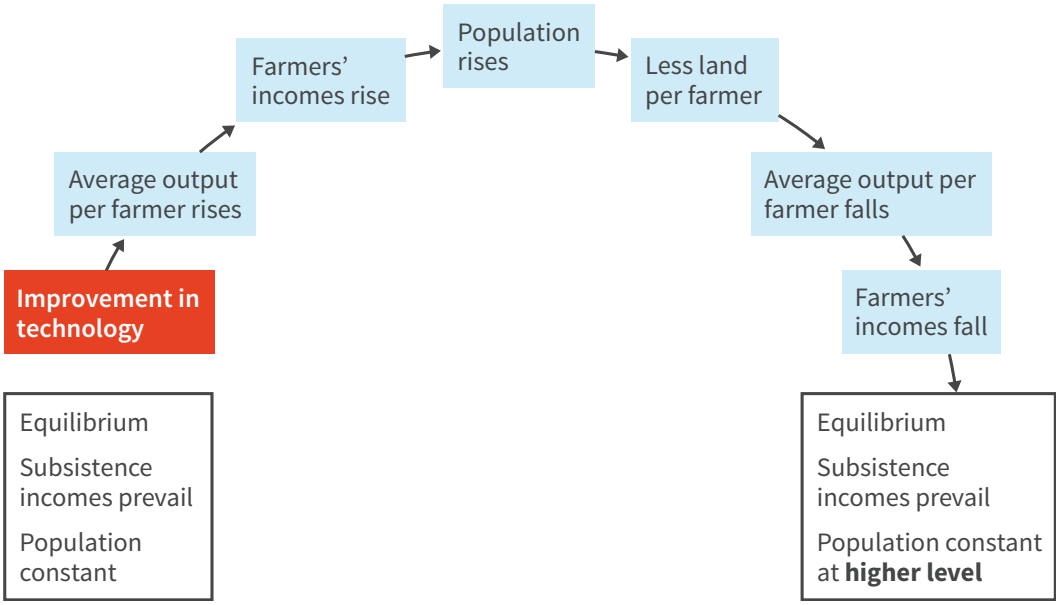


Figure 2.14 *Malthus' model: The effect of an improvement in technology.*

The conditions for the Malthusian model to apply to improvements in technology are:

- The average product of labour diminishes as more labour is applied to a fixed amount of land.
- Population grows in response to increases in real wages.

In this case, in the *long run*, an increase in productivity will result in increased population *but not increased wages*. This depressing conclusion was once regarded as so universal and inescapable that it was called *Malthus' law*.

2.7 THE MALTHUSIAN TRAP AND LONG-TERM ECONOMIC STAGNATION

The major long-run impact of better technology in this Malthusian world was therefore more people, which might explain why China and India, with relatively sophisticated economies at the time, ended up with large populations but—until recently—very low incomes. The writer H. G. Wells, author of *War of the Worlds*, wrote in 1905 that humanity "spent the great gifts of science as rapidly as it got them in a mere insensate multiplication of the common life".

So we now have a possible explanation of the long, flat portion of the hockey stick. Human beings periodically invented better ways of making things, both in agriculture and in industry, and this periodically raised the incomes of farmers and employees above subsistence. The Malthusian interpretation was that higher real wages led young couples to marry earlier and have more children, and they also led to lower death rates. This caused population growth, which eventually forced real wages back to subsistence levels.

As with our model of innovation rents, relative prices and technological improvements, we need to ask: is this Malthusian story consistent with what happened?

Figure 2.15 is consistent with what Malthus predicted. From the end of the 13th century to the beginning of the 17th century Britain oscillated between periods of higher wages, leading to larger populations, leading to lower wages, leading to smaller populations, leading to... and so on, a vicious circle.

We get a different view of the same vicious circle by taking Figure 2.15 and focussing on the period between around 1340 (the beginning of the outbreak of bubonic plague known as the Black Death) and 1600, in Figure 2.16. The lower part of the figure shows the causal linkages that led to the effects we see in the top part.

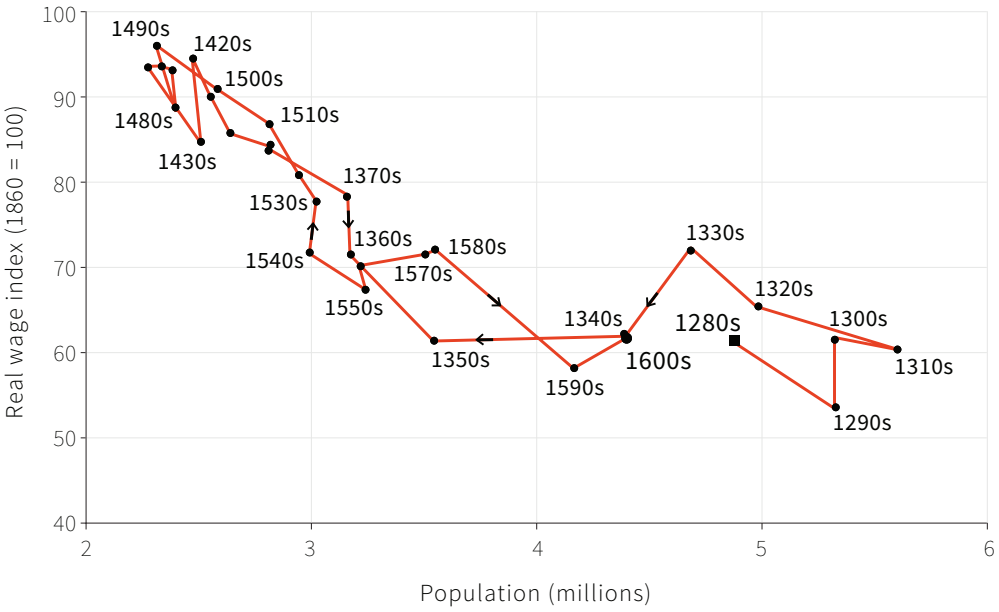


Figure 2.15 *The Malthusian trap: Wages and population (1280-1600).*

Source: *Allen, R. C. (2001), The Great Divergence in European Wages and Prices from the Middle Ages to the First World War.*

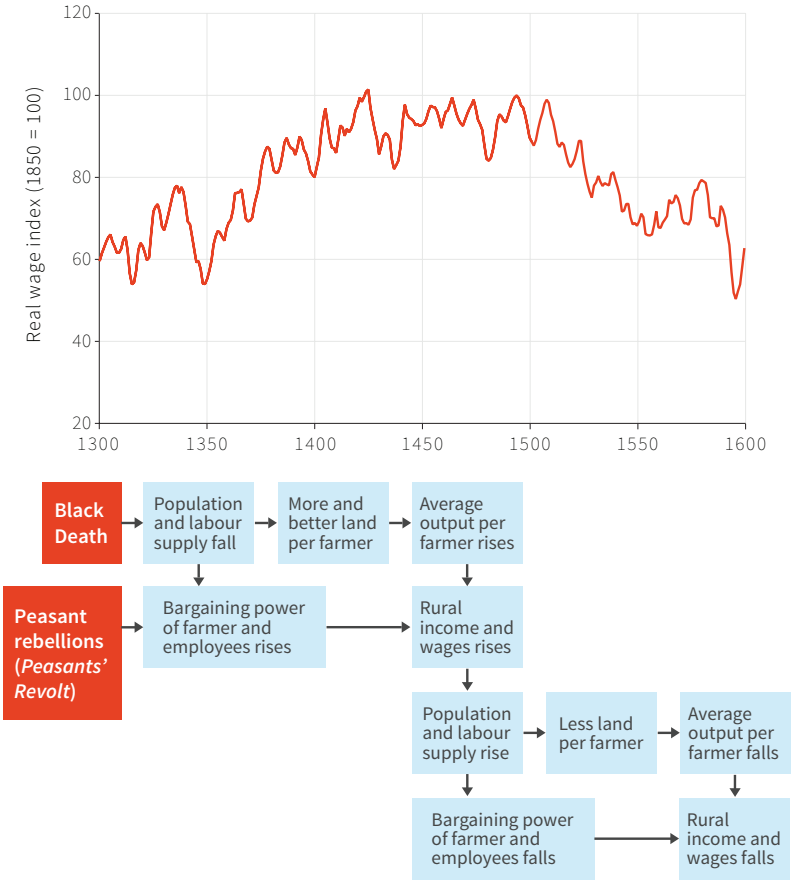


Figure 2.16 *The Black Death, the supply of labour, politics, and the wage: A Malthusian economy.*

The causal links in Figure 2.16 combine the two features of the Malthusian model with the role of political developments as responses to, and causes of changes in, the economy.

In 1349 and 1351, King Edward III of England passed laws to try to restrain wage increases. In this case, economics (the reduced labour supply) won out over politics: wages continued to rise, and peasants began to exercise their increased power, notably by demanding more freedom and lower taxes in a rebellion in 1381.

As a result of the Black Death, between a quarter and a third of Europe's population died between 1349 and 1351. The real wages of English building workers started to rise at the time of the bubonic plague, and had doubled by the middle of the 15th century. In England, the reduction in the number of people working had these effects:

- *An increase in average productivity*: The principle of diminishing average product of labour in farming combined with fewer farmers who had access to the same land.
- *Farmers who owned their land were better off*: They owned what they produced. Farmers paying a fixed rent to a landlord were also better off.
- *Higher wages were offered in cities*: The increased incomes of farmers made it more difficult for employers to attract workers from rural areas.

But, when the population recovered in the 16th century, labour supply increased, which lowered wages. On this evidence, the Malthusian explanation is consistent with the history of England at this time.

DISCUSS 2.7: WHAT WOULD YOU ADD?

The cause-and-effect diagram that we created in Figure 2.16 made use of many ceteris paribus assumptions.

1. In what sense does the model simplify reality?
2. What has been left out?
3. Try redrawing the figure to include other factors that you think are important.

DISCUSS 2.8: DEFINING ECONOMIC PROGRESS

Real wages also rose sharply in other countries for which we have evidence, such as Spain, Italy, Egypt, the Balkans and Constantinople (present-day Istanbul).

"The common people... wanted the dearest and most delicate foods…while children and common women clad themselves in all the fair and costly garments of the illustrious who had died."
Matteo Villani, Florence, Italy (1363)

1. Villani, a resident of Florence, also complained that workers were asking for wages three times higher than before. Why do you think this irritated him?
2. How does the growth of real wages compare with the growth of GDP per capita as a measure of economic progress?
3. What arguments can you propose in favour of each, and what are the drawbacks of each measure?
4. Try out your arguments on others. Do you agree or not? If you disagree, are there any facts that could resolve your disagreement, and what are they? If there are not, why do you disagree?

We have focussed on farmers and wage-earners, but not everyone in the economy would be caught in a Malthusian trap. As population continues to grow, the demand for food also grows; therefore the limited amount of land used to produce the food should become more valuable. In a Malthusian world of rising population, therefore, population growth should lead to an improvement in the relative economic position of landowners.

This occurred in England: the figure shows real wages did not increase in the very long run (they were no higher in 1800 than in 1450). And the gap between landowners and workers increased. In the 17th and 18th centuries the wages of unskilled English workers, relative to the incomes of landowners, they were *only one-fifth of what they been in the 16th century.*

But while wages were low compared to the rents of landlords, this was not the comparison of relative prices that was crucial to how England escaped the Malthusian trap. The key to this process was that wages remained high compared to the price of coal (Figure 2.8) and even increased compared to cost of using capital goods (Figure 2.9), as we have seen.

2.8 ESCAPING FROM MALTHUSIAN STAGNATION

Nassau Senior, the economist who lamented that the numbers perishing in the Irish famine would scarcely be enough to do much good, does not appear compassionate. But he and Malthus were right to think that population growth and a diminishing average product of labour could create a vicious circle of economic stagnation and poverty. The hockey-stick graphs of living standards show they were wrong, however, to believe this could *never* change.

We should revise Malthus' law. An economy would be stuck in the Malthusian trap if it satisfied three, not two, conditions:

- Diminishing average product of labour in production
- Rising population in response to increases in wages
- *An absence of improvements in technology to offset the diminishing average product of labour*

The permanent technological revolution, it turns out, means the Malthusian model is no longer a reasonable description of the world. Average living standards increased rapidly and permanently after the capitalist revolution.

Figure 2.17 shows the real wage and population data from the 1280s to the 1860s. As we saw in Figure 2.15, from the 13th to the 16th century there was a clear negative relationship between population and real wages: when one went up the other went down, just as Malthusian theory suggests.

Between the end of the 16th and the beginning of the 18th century, although wages rose, there was relatively little population growth. Around 1740, we can see the Malthusian relationship again, labelled "18th century". Then, around 1800, the economy moved to what appears to be an entirely new regime, with both population and real wages simultaneously increasing. This is labelled "Escape".

Figure 2.18 zooms in on this "great escape" portion of the wage data.

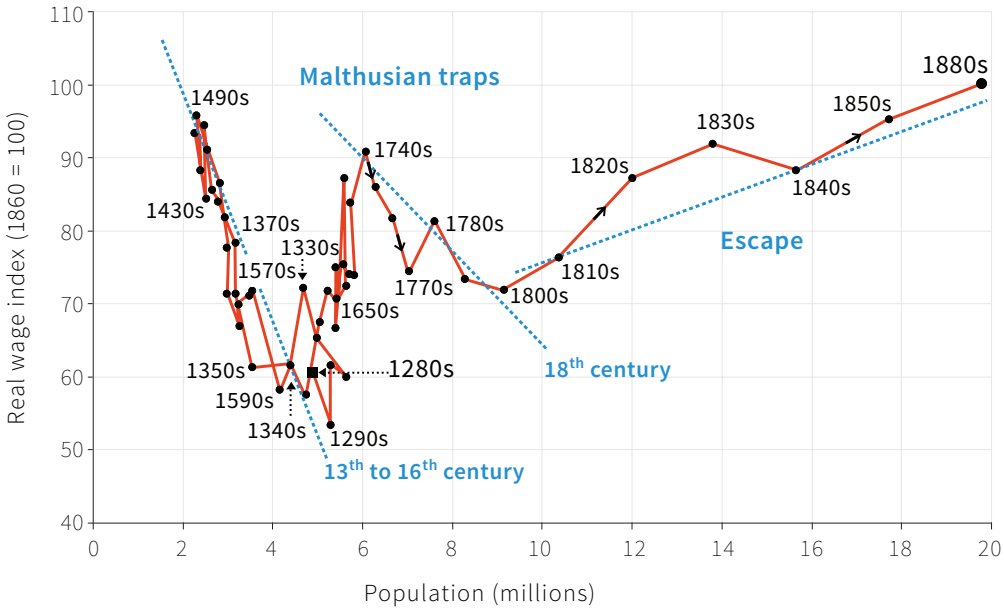


Figure 2.17 *Escaping the Malthusian population trap: Population and real wages in England (1280s-1860s).*

Source: *Allen, R. C. (2001), The Great Divergence in European Wages and Prices from the Middle Ages to the First World War.*

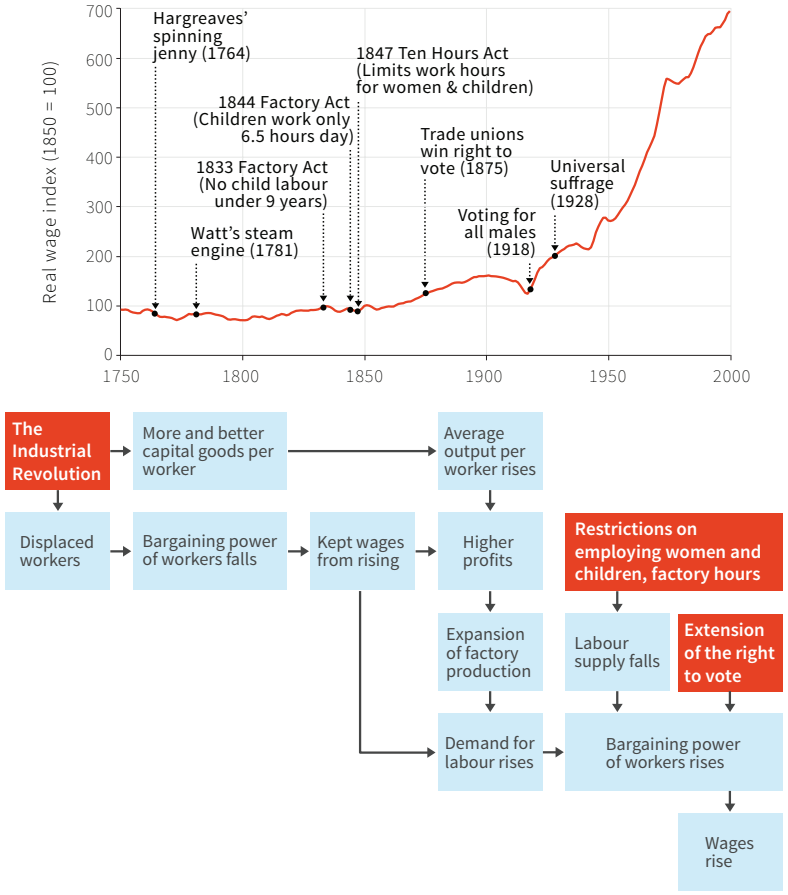


Figure 2.18 *Escaping the Malthusian trap.*

Escaping the Malthusian trap

Figure 2.17 showed that, in the middle of the 18th century, the Malthusian relationship persisted but that the economy moved to what appears to be a new, non-Malthusian regime in the 19th century, with both population and real wages simultaneously increasing. The story begins with technological improvements, increasing output per worker, such as the spinning jenny and the steam engine. Innovations continued throughout the period as the technological revolution became permanent, displacing thousands of spinsters, weavers and farmers.This reduced the bargaining power of workers, keeping wages low, as can be seen by the flat line between 1750 and 1830. During this period the size of the pie was increasing, but the workers' slice was not. In the 1830s the forces of supply and demand for labour led to a surge in profits. Profits, competition and technology made it possible for businesses to expand. The demand for labour went up. People gave up farming, and took jobs in the new factories. The supply of labour went down when business owners were stopped from employing children. The combination of higher demand for, and lower supply of, labour enhanced workers' bargaining power. The power of working people increased as they gained the right to vote and formed trade unions. The result of these changes was that workers were able to claim a constant or even rising share of the increases in productivity that were being generated by the permanent technological revolution.

In one of our Economist in action videos, Suresh Naidu, an economic historian, explains how population growth, technological development and political events interacted to produce the real wage hockey stick.

We see from this dramatic improvement in workers' incomes that there are two influences on wages.

- *How much is produced*: The amount produced by the combination of labour and other inputs establishes a maximum wage level, even if the owner of the land or the machines receives nothing. As in the example of the farming production function in section 2.6, the average product of labour determines the size of the pie to be divided up, but it does not determine how much workers will get.
- *The share going to workers*: This depends on their bargaining power, which in turn depends on how wages are determined (individually, or through bargaining with trade unions, for example) and the supply and demand for workers. If many workers are competing for the same job, wages are likely to be low.

After 1830 the pie continued growing, and the workers' share grew along with it. Britain had escaped from the Malthusian trap. This process would soon be repeated in other countries, as the two Figures 1.1a and 1.1b showed.

WHEN ECONOMISTS DISAGREE

WHAT WERE THE CAUSES OF THE INDUSTRIAL REVOLUTION?

The argument we used to build our model, that Britain adopted technologies before the rest of the world because it was a high-wage and low-energy-cost economy, has been made by Robert Allen, an economic historian. His book Global Economic History: A Very Short Introduction is an accessible primer.

But Allen's explanation of the Industrial Revolution is controversial:

- Joel Mokyr, who has worked extensively on the history of technology, claims that the real sources of technological change are to be found in Europe's scientific revolution and Enlightenment, and in the skilled artisans who made it possible to build the machines of the period. He claims that, while relative factor prices might tilt the direction of invention in one direction or another, they are more like a steering wheel than the motor of technological progress.
- David Landes, a historian, also disagrees with Allen. He suggests Europe pulled ahead of China for cultural and institutional reasons. The Chinese state was too powerful, he argued, and stifled innovation, while Chinese culture favoured stability over change.
- Gregory Clark, an economic historian, also attributes Britain's take-off to culture. But for Clark the keys to success were the cultural attributes such as hard work and savings, which were genetically transmitted to children.
- Kenneth Pomeranz, a historian, claims that superior European growth after 1800 was more due to the abundance of coal in Britain than to any cultural or institutional differences with China. Pomeranz also argues that Britain's access to agricultural production in its New World colonies (especially sugar and its by-products) fed the expanding class of industrial workers without using inferior land, avoiding the diminishing average product of labour problem.

Scholars will probably never completely agree about what caused the Industrial Revolution. One problem is that this change happened only once, which makes it more difficult for social scientists to explain. Also the European takeoff was probably the result of a combination of scientific, demographic, political, geographic and military factors. Several argue that it was partly due to interactions between Europe and the rest of the world too, not just to changes within Europe.

Historians like Pomeranz tend to focus on peculiarities of time and place. They are more likely to conclude that the Industrial Revolution happened because of a unique combination of favourable circumstances (they may disagree about which ones).

Economists like Allen are more likely to look for general mechanisms that can explain success or failure across both time and space.

Economists have much to learn from historians, but often a historian's argument is not precise enough to be testable using a model. On the other hand, historians may regard the economists' models as simplistic, making *ceteris paribus* assumptions that ignore important historical facts. This creative tension is what makes economic history so fascinating.

Economic historians have made progress in recent years in quantifying economic growth over the very long run. By making it clearer what happened, their work makes it easier for us to think about *why* it happened. Some of the work involves comparing real wages in countries over the long run. This has involved collecting both wages and the prices of goods that workers consumed. An even more ambitious series of projects has calculated GDP per capita back to the middle ages.

2.9 CONCLUSION

In this unit:

- We introduced an economic model of how the process of competition among firms both allows the owners of firms that successfully innovate to make extraordinary profits, and also stimulates the diffusion of technological improvements throughout the economy as follower firms try to avoid being left behind.
- We explained how the Malthusian model of the economy created a vicious circle in which population growth devoured temporary gains in income, until the permanent technological revolution allowed an escape due to improvements in technology.
- We explained how the forces of supply and demand, as well as political and other influences on the bargaining power of workers and their employers, help explain this particular hockey stick of history; measuring first stagnation in living standards and then a phenomenal increase.

We have told the story of how the capitalist revolution altered the course of the history of Britain because this was the first occurrence of this unique combination of the new economic system based on private property, markets and firms and the permanent technological revolution.

But it cannot be said that Britain or any of the other countries were typical of this process; each economy that broke out of the Malthusian trap did it by a different escape route. The national trajectories of the early followers were influenced in part by the dominant role that Britain had come to play in the world economy. Germany,

for example, could not compete with Britain in textiles; but the government and large banks played a major role in building steel and other heavy industries. Japan outcompeted even Britain in some Asian textile markets, benefiting from the isolation it enjoyed by the sheer distance from the earlier starters (in those days weeks of travel). Japan selectively copied both technology and institutions, introducing the capitalist economic system while retaining many traditional Japanese institutions including rule by an emperor that would last until the Japanese defeat in the second world war.

CONCEPTS INTRODUCED IN UNIT 2

Before you move on, review these definitions:

- Equilibrium
- Ceteris paribus
- Relative prices
- Incentives
- Diminishing average product of labour
- Reservation option
- Economic rent
- Isocost line
- Innovation rent

India and China provide even greater contrasts. China experienced the capitalist revolution when the Communist Party led a transition away from the centrally planned economy, the antithesis of capitalism that the Party itself had implemented. India by contrast is the first major economy in history to have adopted democracy including universal voting rights prior to its capitalist revolution.

As we saw in Unit 1, the Industrial Revolution did not lead to economic growth everywhere in the world. Because the Industrial Revolution originated in Britain, and spread only slowly to the rest of the world, it also implied a huge increase in income inequality between countries in the 19th and 20th centuries. David Landes, a historian of the Industrial Revolution, once asked "Why are we so rich and they so poor?"

By "we", he meant the rich societies of Europe and North America; by "they" he meant the poorer societies of Africa, Asia and Latin America. Landes suggested, a little mischievously, that there were basically two answers to this question:

"One says that we are so rich and they so poor because we are so good and they so bad; that is, we are hardworking, knowledgeable, educated, well-governed, efficacious, and productive, and they are the reverse. The other says that we are so rich and they so poor because we are so bad and they so good: we are greedy, ruthless, exploitative, aggressive, while they are weak, innocent, virtuous, abused, and vulnerable."
David Landes, *Why are we so rich and they so poor?* (1990)

If you think that the Industrial Revolution happened in Europe because of the Protestant Reformation, or the Renaissance, or the Scientific Revolution, or the development of superior private property rights, or favourable government policies, then you are in the first camp. If you think that it happened because of colonialism, or slavery, or the demands of constant warfare, then you are in the second.

You will notice that these are all non-economic forces that, according to some scholars, had important economic consequences. You can probably also see how the question of which of Landes's two answers is right might become ideologically charged; although, as Landes points out, "It is not clear... that one line of argument necessarily precludes the other".

DISCUSS 2.9: WHY DID THE INDUSTRIAL REVOLUTION NOT HAPPEN IN ASIA?

Read how Landes answered his own question and this analysis of whether Britain's experience was a one-off to discuss why the Industrial Revolution happened in Europe rather than in Asia, and in Britain rather than in Continental Europe.

1. Which arguments do you find most persuasive, and why?
2. Which arguments do you find *least* persuasive, and why?

Key points in Unit 2

Models

Economists use facts and models to understand how the economy works and how it might be made to work better for people.

Relative prices

Relative prices (including wages) influence the choice of technology.

Innovation rents

Innovation rents provide a stimulus for technical innovation.

Competition diffuses technology

Competition creates an environment in which innovations that earn innovation rents are copied, diffusing technological improvements.

Malthusian economics

Malthusian economics uses the principle of diminishing average product of labour to predict that, as the population in a fixed area of land increases, the output obtained by each successive worker will decline. The average output per worker will fall, and living standards will decline as a result.

Subsistence living standards

Population changes in a Malthusian economy, in the long run, drive down living standards to a subsistence level, at which they remain constant.

Escape from the Malthusian trap

By the end of the 19th century rich countries escaped from the trap. Continuous technological progress and a scarcity of labour had created a durable increase in living standards.

Explaining the Industrial Revolution

There are many explanations for why the Industrial Revolution happened in Britain first, but economic incentives in the form of relative prices played a part.

2.10 READ MORE

Bibliography

1. Allen, Robert C. 2011. *Global Economic History: A Very Short Introduction*. New York, NY: Oxford University Press.
2. Allen, Robert C. 2001. 'The Great Divergence in European Wages and Prices from the Middle Ages to the First World War.' *Explorations in Economic History* 38 (4): 411–47.
3. Allen, Robert C. 2009. 'The Industrial Revolution in Miniature: The Spinning Jenny in Britain, France, and India.' *The Journal of Economic History* 69 (04): 901–27.
4. Allen, Robert C. 2008. *The British Industrial Revolution in Global Perspective*. Cambridge: Cambridge University Press.
5. Brainard, William C., and Herbert E. Scarf. 2005. 'How to Compute Equilibrium Prices in 1891.' *American Journal of Economics and Sociology* 64 (1): 57–83.
6. Broadberry, Stephen. 2013. 'Accounting for the Great Divergence.' *VoxEU.org*. November 16.
7. Broadberry, Stephen. 2013. 'Accounting for the Great Divergence.' Economic History Working Papers, 184/13. London School of Economics and Political Science.
8. Clark, Gregory. 2007. *A Farewell to Alms: A Brief Economic History of the World*. Princeton, NJ: Princeton University Press.
9. Davis, Mike. 2001. Late Victorian Holocausts: El Nino Famines and the Making of the Third World. London: Verso Books.
10. Gerschenkron, Alexander. 1962. *Economic Backwardness in Historical Perspective*. Cambridge, MA: Harvard University Press.
11. Greif, Avner, and Guido Tabellini. 2010. 'Cultural and Institutional Bifurcation: China and Europe Compared.' *American Economic Review* 100 (2): 135–40.
12. Herlihy, David. 1997. *The Black Death and the Transformation of the West*. Cambridge, MA: Harvard University Press.
13. Landes, David S. 2006. 'Why Europe and the West? Why Not China?' *Journal of Economic Perspectives* 20 (2): 3–22.
14. Landes, David S. 1990. 'Why Are We So Rich and They So Poor?' *American Economic Review* 80 (May): 1–13.
15. Landes, David S. 2003. *The Unbound Prometheus: Technological Change and Industrial Development in Western Europe from 1750 to the Present*. Cambridge: Cambridge University Press.
16. Lee, James, and Wang Feng. 1999. 'Malthusian Models and Chinese Realities: The Chinese Demographic System 1700-2000.' *Population and Development Review* 25 (1): 33–65.

17. Malthus, Thomas R. 1798. 'An Essay on the Principle of Population.' Library of Economics and Liberty.
18. McNeill, William. 1976. *Plagues and Peoples*. Garden City, NY: Anchor Press.
19. Mokyr, Joel. 2002. *The Gifts of Athena: Historical Origins of the Knowledge Economy*. Princeton, NJ: Princeton University Press.
20. O'Rourke, Kevin H., and Jeffrey G. Williamson. 2005. 'From Malthus to Ohlin: Trade, Industrialisation and Distribution Since 1500.' *Journal of Economic Growth* 10 (1): 5–34.
21. Pomeranz, Kenneth L. 2000. *The Great Divergence: China, Europe and the Making of the Modern World Economy*. Princeton, NJ: Princeton University Press.
22. Schumpeter, Joseph A. 1950. 'The March Into Socialism.' *The American Economic Review* 40 (May): 446–56.
23. Schumpeter, Joseph A. (1942) 2008. *Capitalism, Socialism, and Democracy*. New York, NY: Harper Perennial Modern Thought.
24. Schumpeter, Joseph A. 1949. 'Science and Ideology.' *American Economic Review* 39: 345–59.
25. Schumpeter, Joseph A. (1951) 2004. *Ten Great Economists*. London: Taylor & Francis.
26. Skidelsy, Robert. 2015. 'Portrait: Joseph Schumpeter.' *Skidelskyr.com*. Accessed June.
27. Voigtländer, Nico, and Hans-Joachim Voth. 2013. 'Gifts of Mars: Warfare and Europe's Early Rise to Riches.' *Journal of Economic Perspectives* 27 (4): 165–86.

Beta September 2015 version

3

SCARCITY, WORK AND CHOICE

Shutterstock

HOW INDIVIDUALS DO THE BEST THEY CAN, GIVEN THE CONSTRAINTS THEY FACE, AND HOW THEY RESOLVE THE TRADE-OFF BETWEEN EARNINGS AND FREE TIME

- Decision-making under scarcity is a common problem because we usually have limited means available to meet our objectives
- Economists model these situations: first by defining all of the possible actions
- ... then evaluating which of these actions is best, given the objectives
- Opportunity cost describes an unavoidable trade-off in the presence of scarcity: satisfying one objective more means satisfying other objectives less
- This model can be applied to the question of how much time to spend working, when facing a trade-off between more free time and more income
- This model also helps to explain differences in the hours that people work in different countries and also the changes in our hours of work through history

See www.core-econ.org for the full interactive version of *The Economy* by The CORE Project. Guide yourself through key concepts with clickable figures, test your understanding with multiple choice questions, look up key terms in the glossary, read full mathematical derivations in the Leibniz supplements, watch economists explain their work in Economists in Action – and much more.

Institute for New Economic Thinking
UCL SciencesPo Azim Premji University
FRIENDS PROVIDENT Foundation

Imagine that you are working in New York, in a job that is paying you $15 an hour for a 40-hour working week: so your earnings are $600 per week. There are 24 hours in a day and 168 hours in a week so, after 40 hours of work, you are left with 128 hours of free time for all your non-work activities, including leisure and sleep.

Suppose, by some happy stroke of luck, you are offered a job at a much higher wage—six times higher. Your new hourly wage is $90. Not only that, your prospective employer lets you choose how many hours you work each week.

Will you carry on working 40 hours per week? If you do, your weekly pay will be six times higher than before: $3,600. Or will you decide that you are satisfied with the goods you can buy with your weekly earnings of $600? You can now earn this by cutting your weekly hours to just 6 hours and 40 minutes (a six-day weekend!) If this were your choice, you would enjoy an additional 33 hours and 20 minutes (about 26%) more free time than previously. The Einstein section shows you how we calculated these numbers.

Or would you use the greater hourly wage rate to raise both your weekly earnings and your free time?

The idea of receiving, overnight, a six-fold increase in your hourly wage, and being able to choose your own hours of work, might not seem very realistic. But we know from Unit 2 that technological progress since the Industrial Revolution has been accompanied by a dramatic rise in wages. For example, the average real hourly earnings of American workers increased six-fold during the 20th century. And while employees ordinarily cannot just tell their employer how many hours they want to work, over long periods the typical hours that we work will change. In part, this is a response to how much we prefer to work. As individuals, we can choose part-time work, although this may restrict the jobs open to us. Political parties also respond to the preferences of voters, and changes in typical hours of work have occurred as the result of legislation that, in many countries, imposed maximum working hours.

So have people used economic progress as a way to consume more goods, enjoy more free time, or both? The answer is both, but in different proportions in different countries. While hourly earnings increased by a factor of more than six for 20th century Americans, their average annual work time fell by a little more than one-third. So people at the end of the century enjoyed a four-fold increase in annual earnings with which they could buy goods and services, but a much smaller increase, slightly less than one-fifth, in their free time. (The percentage increase in free time would be higher if you subtracted time spent asleep from free time, but still very small relative to the increase in earnings.) How does this compare with the choice you made when our hypothetical employer offered you a six-fold increase in your wage?

Figure 3.1 shows trends in income and working hours since 1870 in three countries. As in Unit 1, income is measured as per capita GDP in US dollars. This is not the same as average earnings, but gives us a useful indication of average income for

the purpose of comparison across countries and through time. In the late 19th and early 20th century, average income approximately trebled, and hours of work fell substantially. During the rest of the 20th century, income per head rose four-fold. Hours of work continued to fall in the Netherlands and France (albeit more slowly) but levelled off in the US, where there has been little change since 1960.

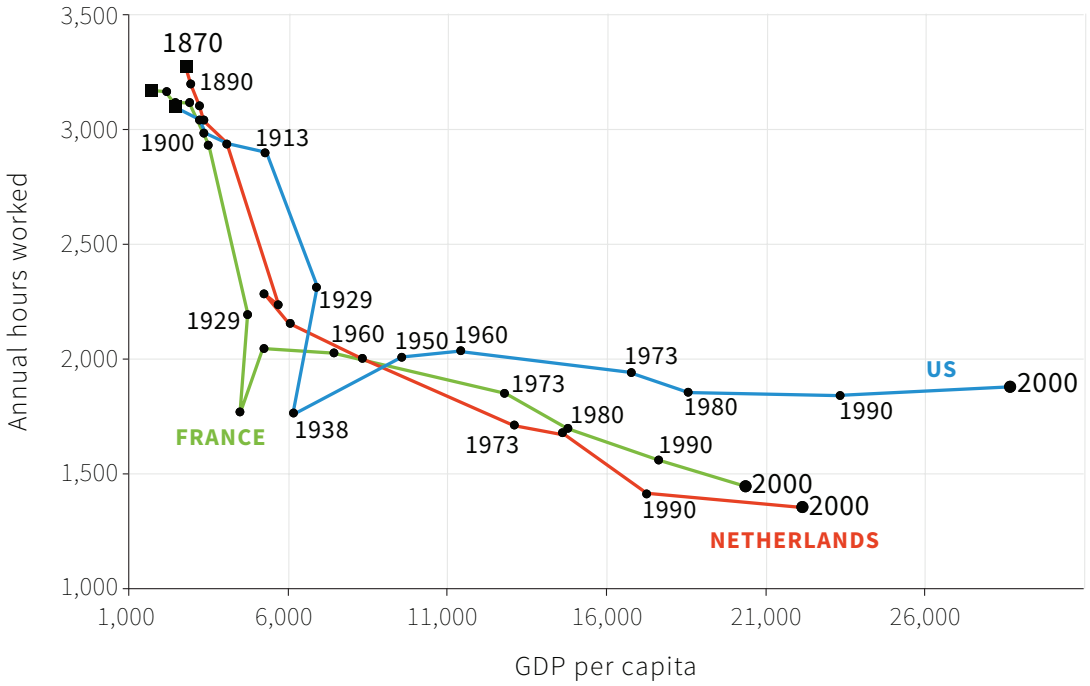


Figure 3.1 *Annual hours of work and income (1870-2000).*

Source: *Maddison Project. 2013. '2013 Edition' . Huberman, Michael, and Chris Minns. 2007. 'The Times They Are Not Changin': Days and Hours of Work in Old and New Worlds, 1870–2000.' Explorations in Economic History 44 (4): 538–67. GDP is measured at PPP in 1990 international Geary-Khamis dollars.*

While many countries have experienced similar trends, Figure 3.2 illustrates the wide disparities between countries in 2013. Here we have calculated free time by subtracting average annual working hours from the number of hours in a year. You can see that the higher-income countries seem to have lower working hours and more free time, but there are some striking differences. For example, the Netherlands and the US have similar levels of income, but Dutch workers have much more free time. And the US and Turkey have similar amounts of free time and a large difference in income.

In many countries there has been a huge increase in living standards since 1870. But in some places people have carried on working just as hard as before, and consumed more, while in other countries people now have much more free time. Why has this happened? We will provide some answers to this question by studying a basic problem of economics—*scarcity*—and how we make choices when we cannot have all of everything that we want, such as goods and free time.

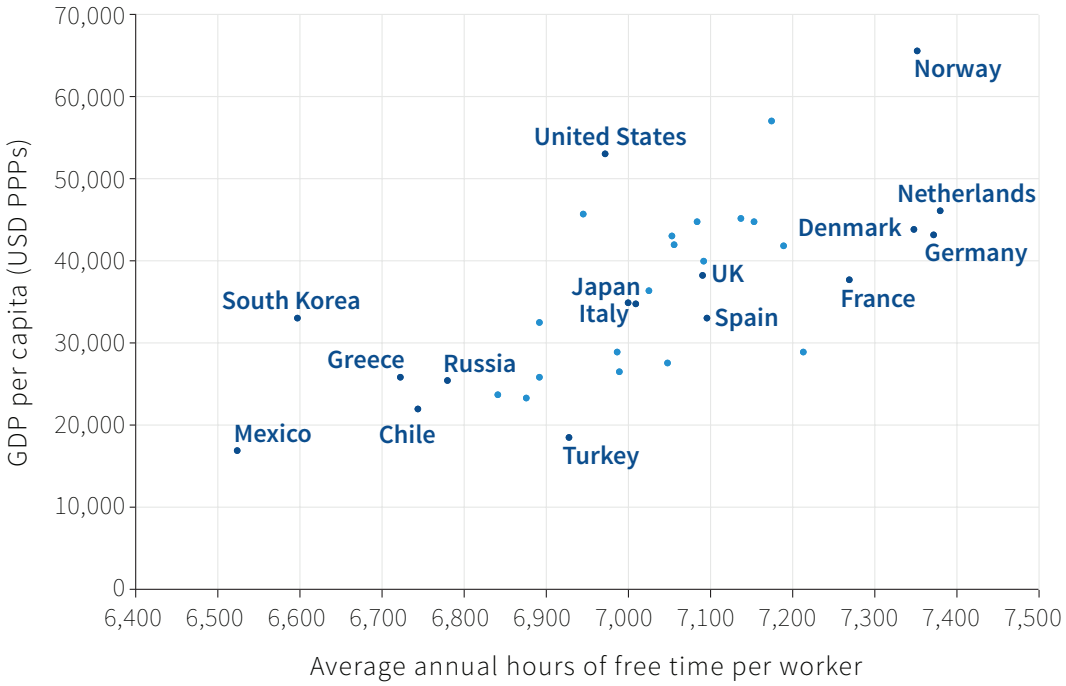


Figure 3.2 *Annual hours of free time per worker and income (2013).*

Source: *OECD. 2015. 'Level of GDP per Capita and Productivity.' Accessed June. OECD. 2015. 'Average Annual Hours Actually Worked per Worker.' Accessed June. Data for South Korea refers to 2012.*

Study the model of decision-making that we use carefully! It will be used repeatedly throughout the course, because it provides insight about a wide range of economic problems.

3.1 LABOUR AND PRODUCTION

In Unit 2 we saw that labour can be thought of as an input in the production of goods and services. Labour is work; in the Unit 1 example of making a cake, it is stirring, mixing, and preparing ingredients. In making a car it is welding, assembling, testing and similar activities. Work activity is often difficult to measure, which will be important in later units because employers find it difficult to determine the amount of work that their employees are doing. We also cannot measure the effort required by different activities in a comparable way (compare baking a cake to building a car), and so economists often measure labour simply as the number of hours worked by individuals engaged in production and assume that, as the number of hours worked increases, the amount of goods produced also increases.

As a student, you make a choice every day: how many hours to spend studying. There may be many factors influencing your choice: how much you enjoy your work, how difficult you find it, how much work your friends do, and so on. Perhaps part of the motivation to devote time to studying comes from your belief that the more time you spend studying, the higher the grade you will be able to obtain at the end of the course. In this unit we will construct a simple model of a student's choice of hours of work, based on the assumption that if you spend more time working, you'll get a better grade.

We assume this is true, but is there any evidence to back this up? A group of educational psychologists looked at the study behaviour of 84 students at Florida State University to identify the factors that affected their performance.

At first sight there seems to be only a weak relationship between the average number of hours per week the students spent studying and their Grade Point Average (GPA) at the end of the semester. This is in Figure 3.3.

	HIGH STUDY TIME (42 STUDENTS)	LOW STUDY TIME (42 STUDENTS)
AVERAGE GPA	3.43	3.36

Figure 3.3 *Study time and grades.*

Source: *Plant, Ashby E., Anders K. Ericsson, Len Hill, and Kia Asberg. 2005. 'Why Study Time Does Not Predict Grade Point Average across College Students: Implications of Deliberate Practice for Academic Performance.' Contemporary Educational Psychology 30 (1): 96–116. Additional calculations were conducted by Ashby Plant, Florida State University, in June 2015.*

The 84 students have been split into two groups according to their hours of study. The average GPA for those with high study time is 3.43—only just above the GPA of those with low study time.

When we look more closely, we discover this is an interesting illustration of why we should be careful when we make *ceteris paribus* assumptions—remember from Unit 2 that this means "holding other things constant". Within each group of 42 students there are many potentially important differences. The conditions in which they study would be an obvious difference to consider: an hour working in a busy, noisy room may not be as useful as an hour spent in the library.

In Figure 3.4, we see that students studying in poor environments are more likely to study for long hours. Perhaps they are distracted by other people around them, so it takes them longer to complete their assignments than students who work in the library. Of these 42 students, 31 of them have high study time, compared with only 11 of the students with good environments.

Now look at the average GPAs in the top row: if the environment is good, students who study longer do better—and you can see in the bottom row that high study time pays off for those who work in poor environments too. This relationship was not as clear when we didn't consider the effect of the study environment.

	HIGH STUDY TIME	LOW STUDY TIME
GOOD ENVIRONMENT	3.63 (11 Students)	3.43 (31 Students)
POOR ENVIRONMENT	3.36 (31 Students)	3.17 (11 Students)

Figure 3.4 *Average GPA in good and poor study environments.*

Source: *Plant, Ashby E., Anders K. Ericsson, Len Hill, and Kia Asberg. 2005. 'Why Study Time Does Not Predict Grade Point Average across College Students: Implications of Deliberate Practice for Academic Performance.' Contemporary Educational Psychology 30 (1): 96–116. Additional calculations were conducted by Ashby Plant, Florida State University, in June 2015.*

So, after taking into account environment and other relevant factors (including the students' past GPAs, and the hours they spent in paid work, or partying) the psychologists estimated that an additional hour of study time per week raised a student's GPA at the end of the semester by 0.24 on average. If we take two students who are the same in all respects except for study time, we predict that the one who studies for longer will have a GPA higher by 0.24 for each extra hour. In other words:

study time raises GPA by 0.24 per hour, ceteris paribus

Now imagine a student, who we will call Alexei, who is able to vary the number of hours he spends studying. We will take the same approach in our model of study time: we assume that *ceteris paribus*, the relationship between the hours Alexei spends studying over the semester and the percentage grade that he will receive at the end is given by the numbers in Figure 3.5. In this model, *study time* refers to all of the time Alexei spends learning, whether in class or individually per day (not per week, as for the Florida State University students). The table shows how the grade will vary if he changes his study hours, if all other factors—his social life, for example—are held constant.

In other words, this is the Alexei's *production function*: it shows how the number of hours per day spent studying (his input of labour) translates into a percentage grade (his output). In reality the grade might also be affected by unpredictable events (in normal life we lump the effect of these things together and call it *luck*). You can think of the production function as telling us what Alexei will get if he is not lucky, but not unlucky either.

If we plot this relationship on a graph, with study time on the horizontal axis and grade on the vertical axis, we get the curve in Figure 3.5. He is able to achieve a higher grade by studying more, so the curve slopes upward. At 15 hours of work per day Alexei gets the highest grade that he is capable of, which is 90%. Any time spent studying beyond that does not affect his exam result (at some point he will be so

tired that studying more each day will not achieve anything), and the curve becomes flat. Work with the interactive figure to see how to calculate his *average product*—the average number of grade point per hour worked—and his *marginal product*—the effect on the grade of studying one more hour.

DISCUSS 3.1: *CETERIS PARIBUS* ASSUMPTIONS

You have been asked to conduct a research study at your university just like the one at Florida State University.

1. What factors do you think should be held constant in a model of the relationship between study hours and final grade?
2. What other information about the students, in addition to study environment, would you want to collect?

Study hours	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15 or more
Grade	0	20	33	42	50	57	63	69	73	78	81	84	86	88	89	90

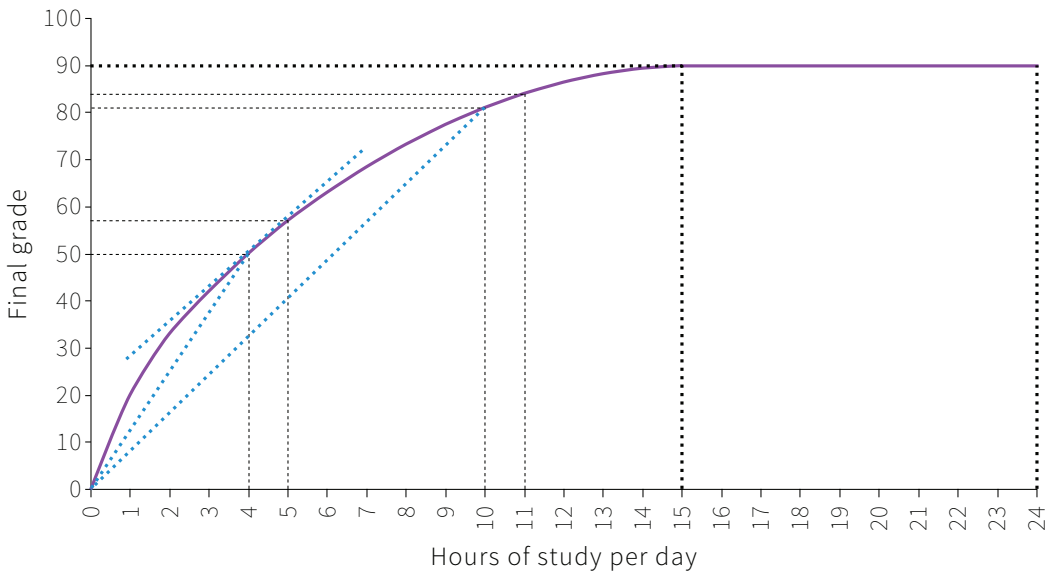


Figure 3.5 *How does the amount of time spent studying affect Alexei's grade?*

Alexei's marginal product is the effect on his grade of studying one more hour. The marginal product corresponds to the slope of the production function. We can see that Alexei's production function in Figure 3.5 gets flatter the more hours he spends studying, so the marginal product of an additional hour per day falls as we move along the curve.

MARGINAL PRODUCT

At each point on the production function, the *marginal product* is the additional amount of output that could be produced if the input was increased by one unit, holding other inputs constant.

The marginal product is *diminishing*. The model captures the idea that extra study helps a lot if you are not studying much; while if you are already studying a lot, then studying even more does not help very much.

If Alexei was already studying for 15 hours a day, he would achieve a final mark of 90. At this combination, what is the marginal product of an additional hour each day? It is zero; studying more does not improve his grade. As you might know from experience, a lack of sleep or of time to relax could even lower Alexei's grade if he worked more than 15 hours a day. If this were the case then his production function would start to slope downward as it approached 24 hours, and Alexei's marginal product would become negative.

In Figure 3.5, output increases as the input increases, but the marginal product falls—the function becomes gradually flatter. A production function with this shape is described as *concave*.

We know that if Alexei works for 4 hours per day he achieves a grade of 50, and his marginal product is 7. We can see the average product—the average number of percentage points Alexei gets per hour of study—in Figure 3.5. It is the slope of a ray from the origin to the curve at 4 hours per day:

$$\begin{aligned} \textit{slope} &= \frac{\textit{vertical distance}}{\textit{horizontal distance}} \\ &= \frac{50}{4} \\ &= 12.5 \end{aligned}$$

Like the marginal product, the average product falls as we move along the curve: at 10 hours of study the ray from the origin is flatter and the average product is only 81/10 =8.1. Why is it smaller? Every time that Alexei decides to work one more hour per day, his production function determines that the *marginal* product of the extra hour of studying will be less than the *average* product of his study without the extra hour. In other words, each *additional* hour of study per day *lowers the average* product of all his study time, taken as a whole.

This is another example of diminishing average product of labour that we saw in Unit 2. In that case the average product of labour in food production (the food produced per worker) may fall as more workers cultivate a fixed area of land.

Marginal change is a common concept, and one that is important in economics. You will often see it marked as a slope on a diagram. With a production function like the one in Figure 3.5, the slope changes continuously as we move along the curve. We

have said that when Alexei studies for 4 hours a day the marginal product is 7, the increase in the grade from one more hour of study. Because the slope of the curve changes between 4 and 5 hours on the horizontal axis, this is an approximation to the marginal product. More precisely, the marginal product is the rate at which the grade increases, per hour of additional study. In Figure 3.5 the true marginal product is the slope of the *tangent* to the curve at 4 hours. In this unit we will use approximations so that we can work in whole numbers, but you may notice that sometimes these numbers are not quite the same as the slopes.

If you know how to use calculus, our Leibniz section shows you how to model the production function algebraically, and find the properties of the average and marginal product of labour.

DISCUSS 3.2: ALEXEI'S PRODUCTION FUNCTION

1. Can you describe a plausible model in which Alexei has a production function that becomes steeper as his hours of work increase?
2. What could cause this to happen?
3. What can you say about the marginal and average products in in this case?

3.2 PREFERENCES

If Alexei has the production function shown in Figure 3.5, how many hours per day will he choose to study? The decision depends on his *preferences*—the things that he cares about. If Alexei cares only about grades, he should study for 15 hours a day. But, in the real world, Alexei also cares about his free time—he likes to sleep, go out or watch TV too. So Alexei faces a trade-off: how many percentage points is he willing to give up in order to do other things when he could be studying?

We illustrate his preferences using Figure 3.6, with *Free time* on the horizontal axis and *Final grade* on the vertical axis. Free time is defined as all the time that he does not spend studying. Every point in the diagram represents a different combination of free time and final grade. Given the production function, not every combination that Alexei would want will be possible, but for the moment we will consider only which combinations Alexei would prefer.

We can assume:

- For a given grade, he prefers a combination with more free time to one with less free time. Therefore, even though both A and B in Figure 3.6 correspond to a grade of 84, Alexei prefers A because it gives him more free time.
- Similarly, if two combinations both have 20 hours of free time, he prefers the one with a higher grade.
- But compare points *A* and *D* in the table. Would Alexei prefer *D* (low grade, plenty of time) or *A* (higher grade, less time)? One way to find out would be to ask him.

Suppose he says he is *indifferent* between *A* and *D*, meaning he would feel equally satisfied with either outcome. We say that these two outcomes would give Alexei the same *utility*. And we know that he prefers *A* to *B*, so *B* provides lower utility than *A* or *D*.

A systematic way to map his preferences would be to start by looking for all of the combinations that give him the same utility as *A* and *D*. We could ask Alexei, another question: "Imagine that you could have the combination at *A* (15 hours of free time, 84 points). How many points would you be willing to sacrifice for an extra hour of free time?" Suppose that—after due consideration—he answers "nine". Then we know that he is indifferent between *A* and E (16 hours, 75 points). Then we could ask the same question about combination *E*, and so on. Eventually we could draw up a table like the one in Figure 3.6. Alexei is indifferent between *A* and *E*, between *E* and *F*, and so on—and that means he is indifferent between *all of these combinations*.

The combinations in the table are plotted in Figure 3.6, and joined together to form a downward-sloping curve, called an indifference curve. The indifference curve joins together all of the combinations that provide equal utility or "satisfaction".

We can draw indifference curves through any point in the diagram, to show other points giving the same utility. If you look at the three curves we drew in Figure 3.6, you can see that the one through *A* gives higher utility than the one through *B*. The curve through *C* gives the lowest utility of the three. To describe preferences we don't need to *measure* an amount of utility; we only need to know which combinations provide more or less of it than others.

The curves we have drawn capture our typical assumptions about people's preferences between two *goods*. In other models, these will often be *consumption goods* such as food or clothing and we refer to the person as a *consumer*. In our model we are looking at the preferences of a student, and the goods are "grade" and "free time". Notice that:

- *Indifference curves slope downward*. If you are indifferent between two combinations, the one that has more of one good must have less of the other good.

- *Higher indifference curves correspond to higher utility levels*. As we move up and to the right in the diagram, further away from the origin, we move to combinations with more of both goods.
- *Indifference curves are usually smooth*. Small changes in the amounts of goods don't cause big jumps in utility.
- *Indifference curves do not cross*. (Why? See Discuss 3.3)
- *As you move to the right along an indifference curve, it becomes flatter*.

	A	*E*	*F*	*G*	*H*	*D*
Hours of free time	15	16	17	18	19	20
Final grade	84	75	67	60	54	50

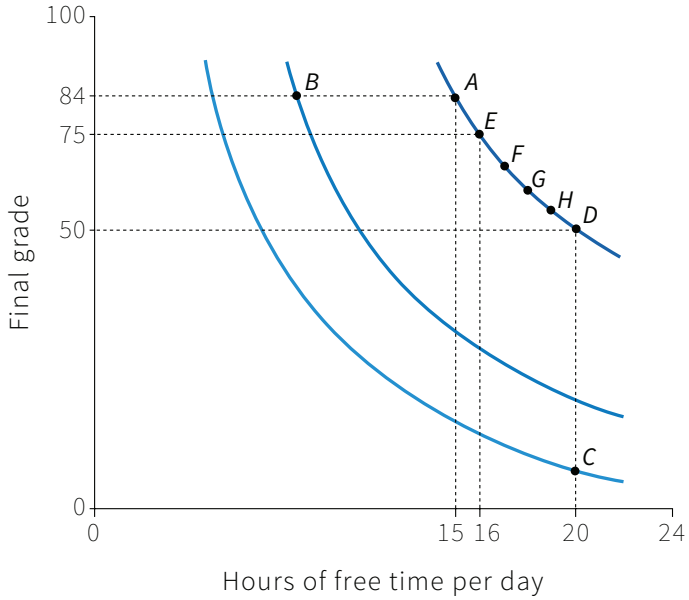


Figure 3.6 *Mapping Alexei's preferences.*

To understand the last property in the list, look at Alexei's indifference curves, which are plotted again in Figure 3.7. If he is at *A*, with 15 hours of free time and a grade of 84, he would be willing to sacrifice 9 percentage points for an extra hour of free time, taking him to *E*: he is indifferent between *A* and *E*. We say that his *marginal rate of substitution* (MRS) between points and free time at *A* is nine; it is the reduction in the grade that would keep Alexei's utility constant following a one-hour increase of free time.

We have drawn the indifference curves as becoming gradually flatter because it seems reasonable to assume that the more free time he has, and the lower the grade, the less willing he will be to sacrifice further percentage points in return for free time; his MRS will be lower. In Figure 3.7 we have calculated the MRS at each combination along the indifference curve. You can see that, when Alexei has more free time and a lower grade, the MRS—the number of percentage points he would give up to get an extra hour of free time—gradually falls.

	A	E	F	G	H	D
Hours of free time	15	16	17	18	19	20
Final grade	84	75	67	60	54	50
Marginal rate of substitution between grade and free time (MRS)	9	8	7	6	4	

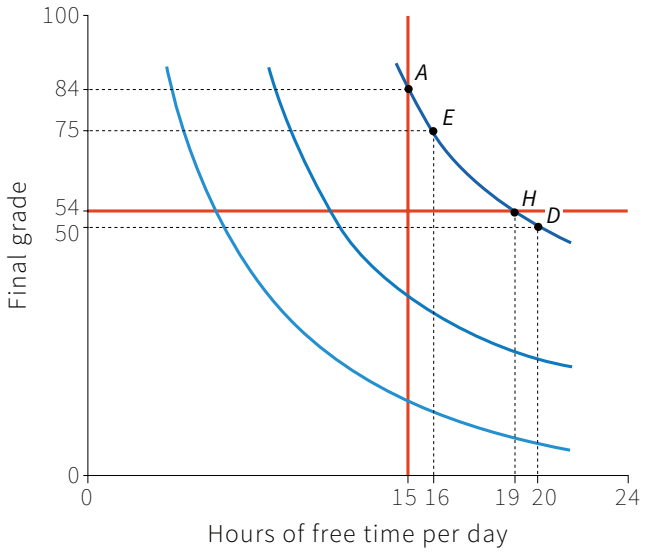


Figure 3.7 *The marginal rate of substitution.*

MARGINAL RATE OF SUBSTITUTION (MRS)

The *MRS* corresponds to the trade-off that a person is willing to make between two goods. At any point, this is the slope of the indifference curve.

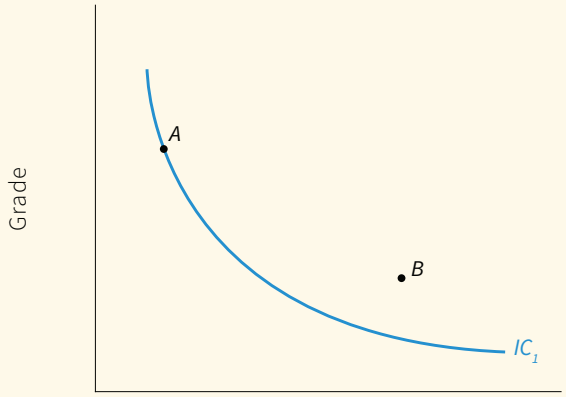
The MRS is just the slope of the indifference curve, and it falls as we move to the right along the indifference curve. In Figure 3.7 you can also see that, for a given amount of free time, Alexei is willing to give up more marks for an additional hour when he has a lot of marks than when he has few (for example, if he was in danger of failing the course). Suppose he has 15 hours of free time. Moving up the vertical line through 15 hours, you can see that the first indifference curve that you meet (the one closest to the origin) is quite flat—the MRS is small. At this combination Alexei's grade is already low: he would not trade many percentage points for an extra hour of time. Moving up to the next indifference curve, where he still has 15 hours of time but a higher grade, the curve is steeper, so the MRS is higher—Alexei is more willing to sacrifice marks because he has more of them. When we reach A, where his grade is 84, the MRS is higher still; marks are so plentiful here that he is willing to give up 9 percentage points for an hour of free time.

You can see the same effect if you fix the grade and vary the amount of free time. If you move to the right along the horizontal line for a grade of 54 the MRS becomes lower at each indifference curve. On the curve closest to the origin, at which the student has least time, the MRS is high. Alexei would be willing to accept a

substantial fall in his grade to have more free time. At *H* he still has 54 percentage points but is much less willing to sacrifice them, because at this point he has plenty (19 hours) of free time already.

DISCUSS 3.3: WHY INDIFFERENCE CURVES NEVER CROSS

In the diagram below, IC_1 is an indifference curve joining all the combinations that give the same level of utility as *A*. Combination B is not on IC_1.



1. Does combination *B* give higher or lower utility than combination *A*? How do you know?
2. Draw a sketch of the diagram, and add another indifference curve, IC_2, through *B* and crossing IC_1. Label the point where they cross as C.
3. Combinations *B* and *C* are both on IC_2. What does that imply about their levels of utility?
4. Combinations *C* and *A* are both on IC_1. What does that imply about their levels of utility?
5. According to your answers to (3) and (4), how do the levels of utility at combinations *A* and *B* compare?
6. Now compare your answers to (1) and (5), and explain how you know that indifference curves can never cross.

DISCUSS 3.4: YOUR MARGINAL RATE OF SUBSTITUTION

Imagine that you are offered at job at the end of your university course that requires you to work for 40 hours per week. This would leave you with 128 hours of free time per week. Estimate the pay that you expect to receive (be realistic!).

1. Draw a diagram with free time on the horizontal axis and pay on the vertical axis, and plot the combination corresponding to your job offer, calling it A. Assume you need about 10 hours a day for sleeping and eating—so it you may want to draw the horizontal axis with 70 hours at the origin.
2. Now imagine you were offered another job requiring 45 hours work per week. What level of pay would make you indifferent between this and the original offer?
3. By asking yourself more questions about the trade-offs you would make, plot an indifference curve through A to represent your preferences.
4. Use your diagram to estimate your marginal rate of substitution between pay and free time at A.

3.3 OPPORTUNITY COSTS

Alexei faces a dilemma: we know from looking at his preferences that he wants both his grade and his free time to be as high as possible, but given his production function, he cannot increase his free time without getting a lower grade in the exam. Another way of expressing this is to say that free time has an *opportunity cost*: to get more free time, Alexei has to forgo the opportunity of getting a higher grade.

OPPORTUNITY COST

When taking an action *A* means forgoing the opportunity of the next best alternative action, *B*, the *opportunity cost* of *A* is the net benefit of action *B*.

In economics opportunity costs are relevant whenever we study individuals choosing between alternative courses of action. In Unit 2 we evaluated a course of action *A* by comparing it with the "next best alternative" action *B*. When we consider the cost of

taking action *A* we include the fact that *if we do A, we cannot do B*. So "not doing *B*" becomes part of the cost of doing *A*. This is called an opportunity cost because doing *A* means forgoing an opportunity to do *B*.

Imagine that an accountant and an economist have been asked to report the cost of going to concert *A*, a concert in a theatre, admission to which costs $25. In a nearby park there is concert *B*, which is free, and happens at the same time.

Accountant The cost of concert *A* is your "out of pocket" cost: you paid $25 for a ticket, so the cost is $25.

Economist But what do you have to give up to go to concert *A*? You gave up $25, *plus the enjoyment of the free concert in the park*. So the cost of the concert for you is the out of pocket cost plus the opportunity cost.

To clarify: suppose that the most you would have been willing to pay to attend the free concert in the park (if it wasn't free) was $15. Then your benefit, were you to choose your next best alternative to concert *A*, would be $15 of enjoyment in the park. This is the opportunity cost of going to concert *A*.

So the total economic cost of concert *A* is *$25 + $15 = $40*. If the pleasure you anticipate from being at concert *A* is $50, then you will forego concert B and buy the ticket to the theatre, because $50 is greater than $40. On the other hand, if you anticipate pleasure from concert *A* of $35, then the economic cost of $40 means you will not choose to go to the theatre. In simple terms, given that you have to pay $25 for the ticket, you will opt for concert *B*—pocketing the $25 to spend on other things and enjoying $15 worth of benefit from the free park concert.

Why don't accountants think this way? Because it is not their job. Accountants are paid to keep track of money, not to provide decision rules on how to choose among alternatives, some of which do not even have a formal price. But making sensible decisions and predicting how sensible people will make decisions involve more than keeping track of money. To see this, we introduce another scenario. Suppose there is no free concert in the park. Your next best alternative is to stay at home, which gives enjoyment of $0:

Accountant Whether or not there is a free park concert does not affect the cost of going to the theatre concert. The cost to you is always $25.

Economist But knowing about the existence of the free park concert helps predict whether you go to concert *A* or not. If your enjoyment from concert *A* is $35 and your next best alternative is staying at home, you will choose concert *A*. However, if concert *B* is available, you will choose it over concert *A*.

In Unit 2, we said that if an action brings greater net benefits than the next best alternative, it yields an *economic rent* and you will do it. Another way of saying this is that you receive an economic rent from taking an action when it results in a benefit greater than its economic cost (that is, both out of pocket and opportunity costs).

Figure 3.8 summarises the example of your choice of which concert to attend.

	A HIGH VALUE ON THE THEATRE CHOICE (A)	A LOW VALUE ON THE THEATRE CHOICE (A)
OUT OF POCKET COST (PRICE OF TICKET FOR A)	$25	$25
OPPORTUNITY COST (FOREGONE PLEASURE OF B, PARK CONCERT)	$15	$15
ECONOMIC COST (SUM OF OUT OF POCKET AND OPPORTUNITY COST)	$40	$40
ENJOYMENT OF THEATRE CONCERT (A)	$50	$35
ECONOMIC RENT (ENJOYMENT MINUS ECONOMIC COST)	$10	-$5
Decision:	*A*: Go to the theatre concert.	*B*: Go to the park concert.

Figure 3.8 *Which concert will you choose? Opportunity costs and economic rent.*

DISCUSS 3.5: OPPORTUNITY COSTS

The British government introduced legislation in 2012 so that universities had the option to raise their tuition fees. Most chose to increase annual tuition fees for students from £3,000 to £9,000.

Does this mean that the cost of going to university has tripled? (Think about how an accountant and an economist might answer this question.)

3.4 THE FEASIBLE SET

Now we return to Alexei's problem of how to choose between grades and free time. We have shown that free time has an opportunity cost in the form of lost percentage points in his grade (equivalently, we might say that percentage points have an opportunity cost in the form of the free time Alexei has to give up to obtain them). But before we can describe how Alexei resolves his dilemma, we need to work out precisely which alternatives are available to him.

To answer this question, it helps to look again at the production function. This time we will show the relationship between final grade and free time, rather than between final grade and study time. There are 24 hours in a day. Alexei must divide this time between studying (all of the hours devoted to learning) and free time (all the rest of his time). Figure 3.9 shows the relationship between his final grade and hours of free time per day—the mirror image of Figure 3.5. If Alexei studies solidly for 24 hours, that means zero hours of free time and a final grade of 90. If he chooses 24 hours of free time per day, we assume he will get no marks at all.

In Figure 3.9, the axes are final grade and free time, the two goods that give Alexei utility. If we think of him choosing to consume a combination of these two goods, the curved line in Figure 3.9 represents his *feasible frontier*. It plots the highest grade he can achieve given the amount of free time he takes. Work with the interactive figure to see which combinations of grade and free time are feasible, and which are not.

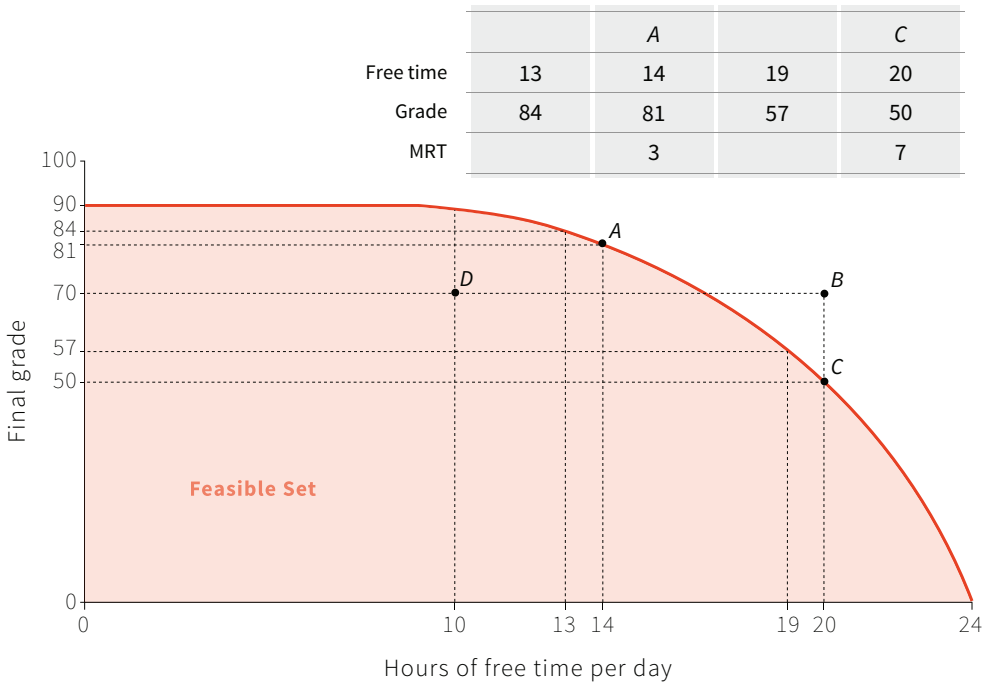


Figure 3.9 *How does Alexei's choice of free time affect his grade?*

Any combination of free time and final grade lying on the frontier is feasible. Combinations outside the feasible frontier are said to be *infeasible* given his abilities and conditions of study. On the other hand a combination lying inside the frontier is feasible; but it would imply Alexei has effectively thrown away something that he values. If he studied for 14 hours a day then, according to our model, he could guarantee himself a grade of 89. But he could obtain a lower grade (70, say), if he just stopped writing before the end of the exam. It would be foolish to throw away marks like this for no reason—but it would be possible. Another way to obtain a combination inside the frontier would be to sit in the library doing nothing—Alexei would be taking less free time than is available to him, which again makes no sense.

By choosing a combination inside the frontier, Alexei would be giving up something that is freely available—something that has no opportunity cost. At combinations inside the frontier he can obtain a higher grade without sacrificing any free time, or have more time without reducing his grade.

The area inside the frontier, together with the frontier itself, is called the *feasible set*; it shows all the combinations of grade and free time per day that are obtainable, *ceteris paribus*.

The feasible frontier is a constraint on the choices Alexei can make. It represents the trade-off he must make between grade and free time. Remember that the slope of the production function corresponds to the marginal product of an hour of study (how many percentage points one more hour of study per day produces), and that the marginal product diminishes as hours of study increase. So, in Figure 3.9, the feasible frontier gets steeper as hours of free time increase (that is, as we move along the frontier from left to right).

The slope of the feasible frontier corresponds to the *marginal rate of transformation* (MRT) between free time and percentage points in the final grade. Alexei can "transform" free time into a higher grade, and the MRT is the number of percentage points he would gain in return for giving up (transforming) one more hour. If he has 14 hours of free time, he could increase his grade from 81 to 84 by giving up one more hour: so his MRT is three. Go back to the last three steps of Figure 3.9 to see how the MRT changes as we move along the feasible frontier. The more free time he takes, the higher the marginal product of studying, so the MRT is higher and the frontier is steeper.

MARGINAL RATE OF TRANSFORMATION (MRT)

The *MRT* is the quantity of some good that must be sacrificed to acquire one additional unit of another good. At any point, it is the slope of the feasible frontier.

Note that we have now identified two trade-offs:

- *The marginal rate of substitution (MRS)*: In the previous section, we saw that it measures the trade-off that the student is willing to make between exam marks and free time.
- *The marginal rate of transformation (MRT)*: In contrast, this measures the trade-off that the student is constrained to make by the feasible frontier.

This Leibniz shows you how to find the find the MRS and MRT using calculus.

As we shall see in the next section, the choice Alexei makes between his grade and his free time will balance these two trade-offs against each other.

3.5 DECISION-MAKING AND SCARCITY

The final step is to look for the combination of grade and free time that Alexei will choose. Figure 3.10 brings together his feasible frontier (Figure 3.9) and indifference curves (Figure 3.6). Therefore it shows both the constraint trade-off and the preference trade-off. Using the figure, we can draw conclusions about the decisions Alexei will make. Recall that the indifference curves indicate what Alexei prefers, and the feasible frontier is the constraint on his choice.

Figure 3.10a shows four indifference curves, labelled IC_1 to IC_4. IC_4 represents the highest level of utility because it is the furthest away from the origin. No combinations of grade and free time on IC_4 are feasible; the whole indifference curve lies outside the feasible set. Suppose that Alexei considers choosing a combination somewhere in the feasible set, on IC_1. By working through the steps in Figure 3.10a you will see that he can increase his utility by moving to points on higher indifference curves, until he reaches a feasible choice that maximises his utility.

Alexei maximises his utility at point E, at which his indifference curve is tangent to the feasible frontier. This example of constrained choice over free time and study (and hence final grade) tells us that, under the assumptions we have made, Alexei will:

- Choose to spend 5 hours each day studying
- Spend 19 hours each day doing other activities
- Obtain a grade of 57 as a result

This choice maximises Alexei's utility.

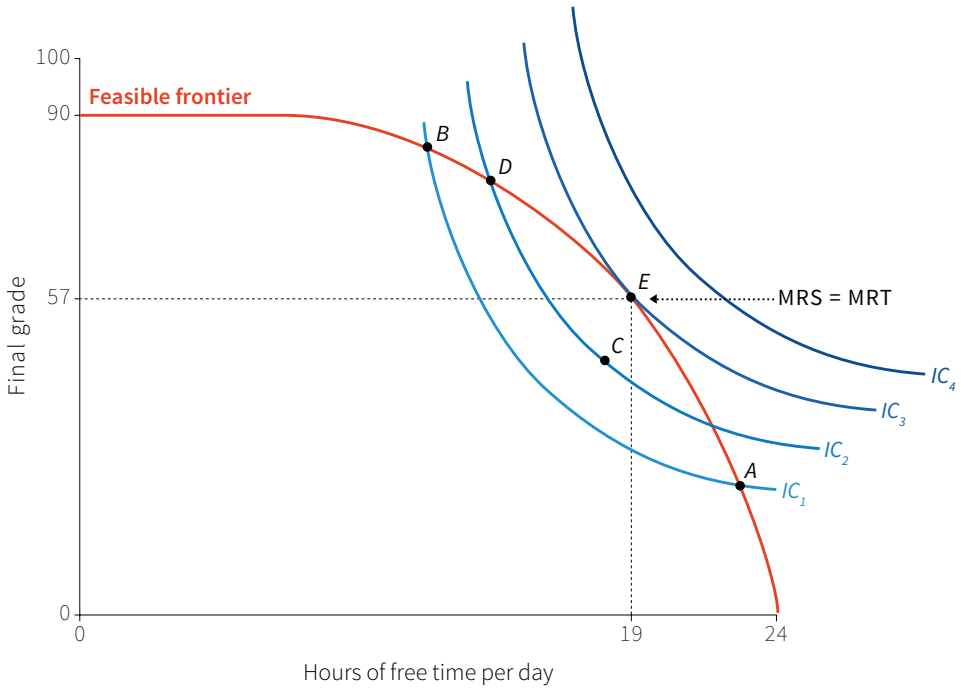


Figure 3.10a *How many hours does Alexei decide to study?*

We can see from Figure 3.10a that, at *E*, the feasible frontier and the highest attainable indifference curve IC_3 are tangent to each other (they touch but do not cross). At *E* the slope of the indifference curve is the same as the slope of the feasible frontier. Now, remember that the slopes represent the two trade-offs facing Alexei:

- *The slope of the indifference curve is the MRS*, which is the trade-off he is willing to make between free time and percentage points.
- *The slope of the frontier is the MRT*, the trade-off that he is constrained to make because it is not possible to go beyond the feasible frontier.

Therefore Alexei achieves the highest possible utility where the two trade-offs just balance (*E*). Alexei's optimal combination of grade and free time is at the point that the marginal rate of transformation is equal to the marginal rate of substitution.

To find out how to determine point E using calculus, see this Leibniz.

Figure 3.10b shows the MRS and MRT at the points shown in Figure 3.10a on the feasible frontier. The MRT is calculated by looking at what happens when Alexei gives up one hour of free time. We have measured the MRS from the slope of the indifference curve. At *B* and *D*, the number of points he is *willing* to trade for an hour of free time (MRS) is greater than the opportunity cost of that hour (MRT) so he prefers to increase his free time. At *A* the MRT is greater than the MRS so he prefers to decrease his free time. And, as expected, at *E* the MRS and MRT are equal.

		B		D		E		A
FREE TIME	12	13	14	15	18	19	21	22
GRADE	86	84	81	78	63	57	42	33
MRT		2		4		7		9
MRS		20		15		7		3

Figure 3.10b *How many hours does Alexei decide to study?*

We have modelled the student's decision on study hours as what we call a *constrained choice problem*: a decision-maker (Alexei) pursues an objective (in this case to maximise his utility) subject to a constraint (his feasible frontier).

In our example, both free time and points in the exam are scarce for Alexei because:

1. *Free time and grades are both goods*: Alexei values both of them.
2. *Each has an opportunity cost*: More of one means less of the other.

CONSTRAINED CHOICE PROBLEMS

These problems provide a way to think rigorously about how to do the best for ourselves:

- Given our preferences
- Given the constraints we face
- ... when the things we value are scarce

In such problems, the solution of the constrained choice problem is the individual's optimal choice. If we assume that utility maximisation is Alexei's goal, *the optimal combination of grade and free time is a point on the feasible frontier at which:*

$$MRS = MRT$$

DISCUSS 3.6: EXPLORING SCARCITY

Describe a situation in which the student's grade points and free time would not be scarce. Remember scarcity depends on both the student's preferences and the production function.

Figure 3.11 summarises Alexei's trade-offs:

	THE TRADE-OFF	WHERE IT IS ON THE DIAGRAM	IT IS EQUAL TO...
MRS	*Marginal rate of substitution*: The number of percentage points Alexei is willing to trade for an hour of free time.	The slope of the indifference curve.	
MRT, OR OPPORTUNITY COST OF FREE TIME	*Marginal rate of transformation*: The number of percentage points Alexei would gain (or lose) by giving up (or taking) another hour of free time.	The slope of the feasible frontier.	The marginal product of labour.

Figure 3.11 *Alexei's trade-offs.*

3.6 HOURS OF WORK AND ECONOMIC GROWTH

One of the aims of this unit is to examine how living standards might change as a result of the choices people make in response to technological progress. As we saw in Unit 2, new technologies raise the productivity of labour. We now have the tools to analyse the effects of increased productivity on living standards: specifically, on both the incomes and the free time of workers. More generally, these tools (built on concepts relating to opportunity costs and preferences) will prove useful in a variety of situations in which choices have to be made in conditions of scarcity.

So far, we have looked at Alexei, a student, and his choice between studying and free time. We now use our model of constrained choice to look at Angela, a self-sufficient farmer who chooses how many hours to work. We assume that Angela produces grain to eat and does not sell it to anyone else. If she produces too little grain, she will starve.

What is stopping her producing the most grain possible? Just like the student, Angela also values free time—so she gets utility from both free time and consuming grain.

But her choice is constrained: grain can be consumed only if it is produced, production takes labour time, and each hour of labour means Angela foregoes an hour of free time. The hour of free time sacrificed is the opportunity cost of the grain produced. Like Alexei, Angela faces a problem of scarcity: she has to make a choice between her consumption of grain and her consumption of free time.

We are interested in two questions:

- *How many hours will Angela choose to work*, given the initial production function?
- *Imagine the production function changes*: an improvement in technology means Angela can produce the same amount of grain with fewer hours of work. *How much more free time will she choose?*

We begin by considering the relationships in Figures 3.12 and 3.13.

Figure 3.12 shows the production function for the initial technology. The table shows how the number of hours Angela works per day affects the amount of grain produced. We can see that the corresponding graph has a similar shape to the student's production function: the marginal product of an additional hour's work is diminishing as we move along the curve, and so is the average product (grain produced per hour).

Working hours	0	1	2	3	4	5	6	7	8	9	10	11	12	14	18	24
Grain	0	9	18	26	33	40	46	51	55	58	60	62	64	66	69	72

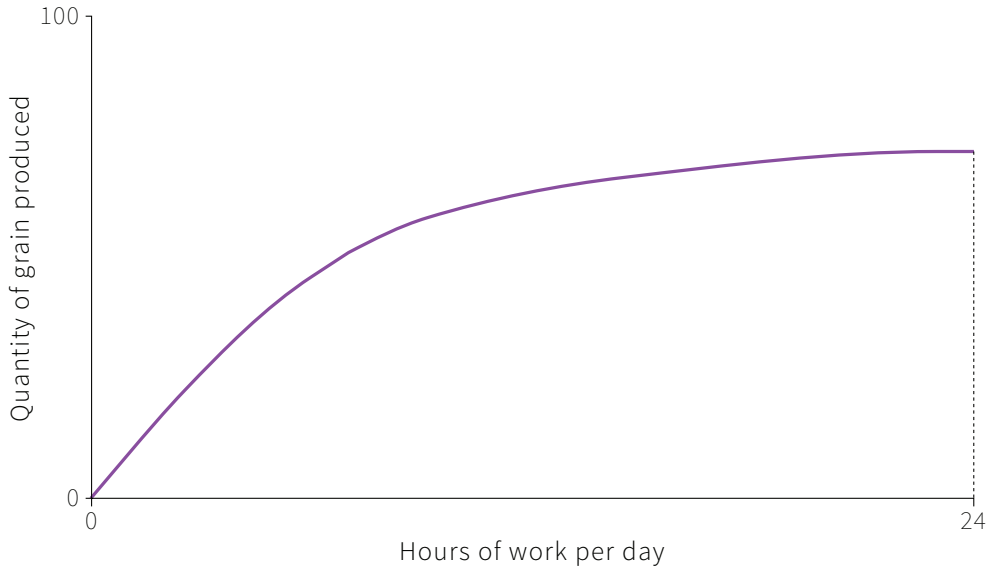


Figure 3.12 *A grain technology: Angela's production function.*

Figure 3.13 shows Angela's feasible frontier, which is just the mirror image of the production function. As before, what we call free time is all of the time that is not spent working to produce grain—it includes time for eating, sleeping, and everything else that we don't count as farm work, as well as her leisure. The diagram shows how much grain can be consumed for each possible amount of free time, given the initial technology. Reflecting the diminishing marginal product of labour in Angela's production function, the *feasible frontier* gets steeper as hours of free time increase: the marginal rate of transformation (MRT) between free time and quantity of grain produced increases as we move along the curve. In simple terms, this means the additional amount of grain that can be produced from giving up an hour of free time is higher when Angela has a lot of free time already.

By bringing together Angela's indifference curves with her feasible set in Figure 3.13, we can find her optimal choice of free time and grain—the feasible combination that maximises her utility.

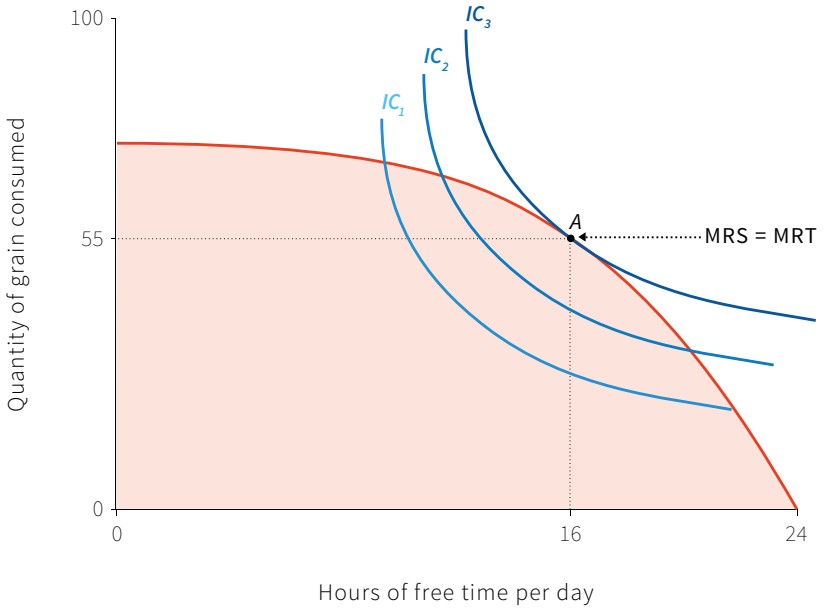


Figure 3.13 *Angela's choice between free time and grain.*

The highest indifference curve Angela can attain, given the feasible frontier, is IC_3. Angela will therefore maximise her utility at point A, enjoying 16 hours of free time per day and consuming 55 units of grain. Just like the student, Angela is balancing two trade-offs at this point: her marginal rate of substitution (MRS) between grain and free time (the slope of the indifference curve) is equal to the MRT (the slope of the feasible frontier). We can think of the combination of free time and grain at point A as representing her standard of living.

Next, we want to think about how Angela's choice of free time and grain responds to an improvement in technology. A technological improvement will increase the amount of grain Angela can produce in a given number of hours of work. This improvement could be better seeds that yield more grain, or better equipment that makes harvesting quicker.

In our model of constrained choice, an improvement in technology shifts the production function upward as shown in Figure 3.14. As before, we abbreviate the labels, denoting the initial production function as PF and the production function after the improvement in technology as PF_{new}.

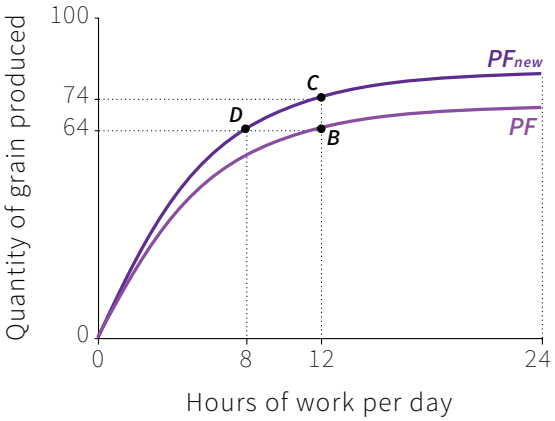


Figure 3.14 *Angela's production function after an improvement in technology.*

If Angela worked for 12 hours per day before the introduction of the new technology, then she would be at *B* on the production function and would be able to produce 64 units of grain. After the technological improvement, Angela is able to produce 74 units of grain by working for 12 hours (point *C*). Alternatively, reducing the hours of work to just eight per day still produces the 64 units of grain she was producing before the improvement in technology (point *D*).

To find out how to model technical change algebraically read this Leibniz.

The new technology has therefore given Angela the option of a number of combinations of free time and grain that were not previously available. In our model this means an upward shift in the feasible frontier as shown in Figure 3.15. We label the initial feasible frontier as *FF* and the feasible frontier after the improvement in technology as FF_{new}. The figure is a mirror image of Figure 3.14; *B*, *C* and *D* in the two figures represent exactly the same combinations of free time and grain.

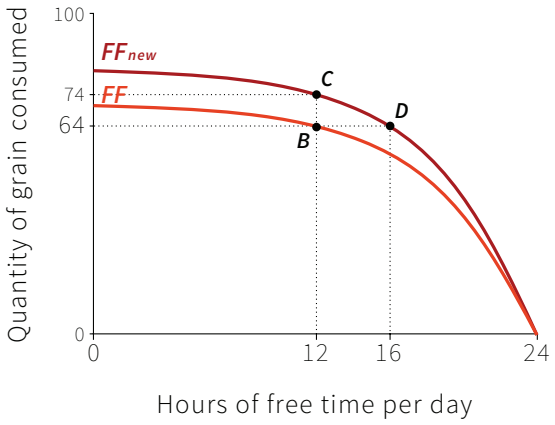


Figure 3.15 *Angela's feasible frontier after an improvement in technology.*

What combination of free time and grain will Angela choose after the improvement in technology? To answer this question, we return to her choice before the technological improvement. As shown by point A in Figure 3.16, she is enjoying 16 hours of free time per day and is consuming 55 units of grain. The improvement in technology shifts the feasible set upward to FF_{new}. Angela can now reach an indifference curve further from the origin because the feasible set has expanded. Figure 3.16 demonstrates that she moves to a point on a higher indifference curve, increasing both her consumption of grain and her free time.

The result is that Angela has responded to the technological improvement by taking some additional free time and consuming more grain. It is important to realise that this is just one possible result. Had we drawn the indifference curves or the frontier differently, the trade-off would have been different. We can say definitely that the improvement in technology makes it *feasible* to both consume more grain and have more free time, but whether Angela will choose to have more of both depends on her preferences between the two goods, and her willingness to substitute one for the other.

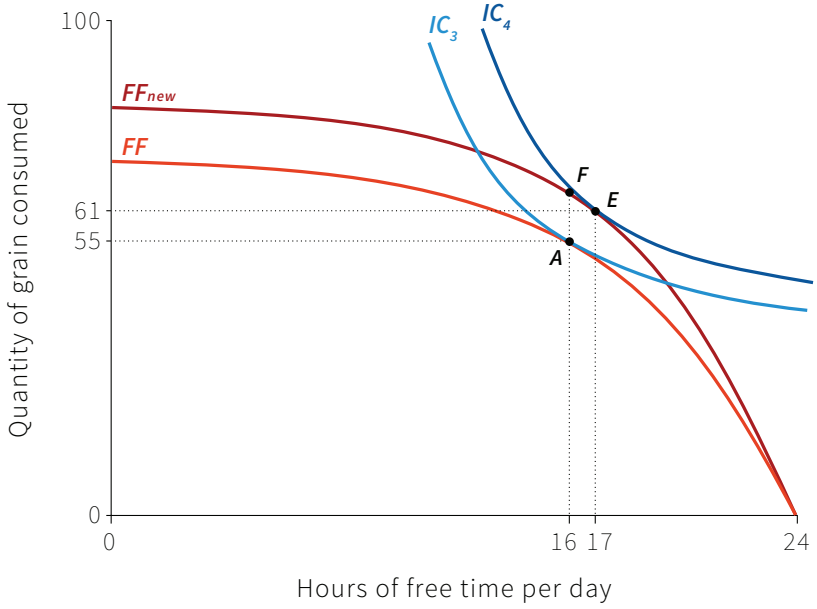


Figure 3.16 *Angela's choice between free time and grain after an improvement in technology.*

To see why look at the last step in Figure 3.16. Notice that first the technological shift results in the feasible frontier becoming steeper at each value of free time: the MRT is higher at F than A. In other words, in the trade-off that has to be made between grain and free time, each additional hour of free time incurs a greater opportunity cost in forgone consumption of grain than was the case prior to the improvement in technology. Taken on its own, this means that the technological improvement generates an increased incentive for Angela to work. But secondly, the expansion in the feasible set means that she can produce more grain for the same amount of work and Angela's indifference curves get steeper as the amount of grain increases—her MRS is higher. So she is now more willing to sacrifice grain for some extra free time. This effect works in the opposite direction to the first one—she has a stronger preference for free time. In Figure 3.16, the second effect dominates and she chooses point *E*, with more free time as well as more grain.

Now suppose that her MRS doesn't change much as her consumption of grain rises. You can see a case like this in Figure 3.17—the slopes of the indifference curves stay the same as we move up the vertical line at 16 hours. Although she can now have

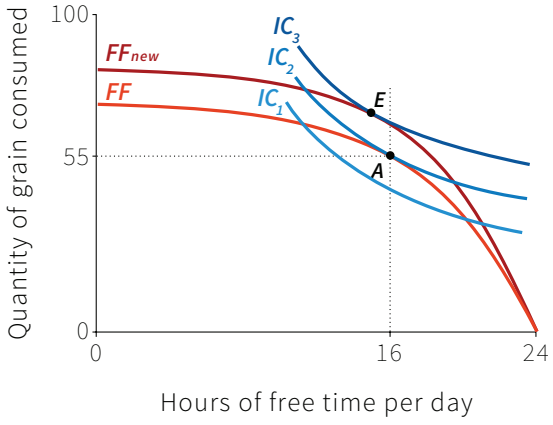


Figure 3.17 *Angela's response to an improvement in technology, with different preferences.*

more grain, she is just as willing as before to substitute between grain and free time. But the opportunity cost of free time is now greater, so she takes less of it, choosing point *E*.

The model of the self-sufficient farmer shows us that technological improvement can lead to Angela working more hours, or working fewer hours. She faces both a stronger incentive to work, because each hour is more productive, and an increased desire for free time because she has more grain. In the next section we look more carefully at these two opposing effects, using a different example to disentangle them.

DISCUSS 3.7: YOUR PRODUCTION FUNCTION

1. What could bring about a technological improvement in the production functions of you and your fellow students?
2. Draw a diagram to illustrate how this improvement would affect your feasible set of grades and study hours.
3. Analyse what might happen to your choice of study hours, and the choices that your colleagues might make.

3.7 INCOME AND SUBSTITUTION EFFECTS ON HOURS OF WORK AND FREE TIME

Imagine that you are looking for a job after you leave college. You expect to be able to earn a wage of $15 per hour. Jobs differ according to the number of hours you work—so what would be your ideal number of hours? Together, the wage and the hours of work will determine how much free time you will have, and your total earnings.

As for Angela, we will work in terms of daily average free time and consumption. We will assume that your spending—that is, your average consumption of food, accommodation, and other goods and services—cannot exceed your earnings (for example, you will not borrow to increase your consumption). If we write w for the wage, and you have t hours of free time per day, then you work for $(24\text{-}t)$ hours, and your maximum level of consumption, c, is given by:

$$c = w(24 - t)$$

We will call this your *budget constraint*, because it shows what you can afford to buy. In the table in Figure 3.18 we have calculated your free time for hours of work varying between 0 and 16 hours per day, and your maximum consumption, when your wage is $w = \$15$.

Figure 3.18 shows the two goods in this problem: *Hours of free time* on the horizontal axis, and *Consumption* on the vertical axis. When we plot the points shown in the table we get a downward-sloping straight line: this is the graph of the budget constraint. The equation of the budget constraint is:

$$c = 15(24 - t)$$

The slope of the budget constraint corresponds to the wage: for each additional hour of free time, consumption must decrease by \$15. The area under the budget constraint is your feasible set; your problem is quite similar to Angela's problem, except that your feasible frontier is a straight line. Remember that for Angela the slope of the feasible frontier is both the MRT (the rate at which free time could be transformed into grain) and the opportunity cost of an hour of free time (the grain foregone). These vary because Angela's marginal product changes with her hours of work. For you, the marginal rate at which you can transform free time into consumption, and the opportunity cost of free time, is equal to your wage: it is \$15 for your first hour of work, and still \$15 for every hour after that.

What would be your ideal job? Your preferred choice of free time and consumption will be the combination on the feasible frontier that is on the highest possible indifference curve. Work through Figure 3.18 to find the optimal choice.

Hours of work	0	2	4	6	8	10	12	14	16
Free time, t	24	22	20	18	16	14	12	10	8
Consumption, c (\$)	0	30	60	90	120	150	180	210	240

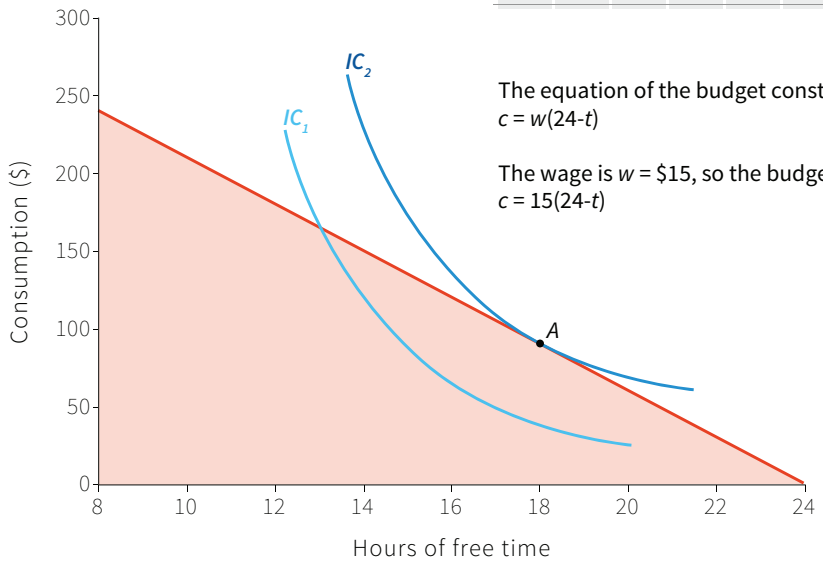


Figure 3.18 *Your preferred choice of free time and consumption.*

If your indifference curves look like the ones in Figure 3.18, then you would choose point A, with 18 hours of free time. At this point your MRS—the rate at which you are willing to swap consumption for time—is equal to the wage ($15, the opportunity cost of time). You would like to find a job in which you can work for 6 hours per day, and your daily earnings would be $90.

Like the student, you are balancing two trade-offs:

	THE TRADE-OFF	WHERE IT IS ON THE DIAGRAM
MRS	*Marginal rate of substitution*: The amount of consumption you are willing to trade for an hour of free time.	The slope of the indifference curve.
MRT, OR OPPORTUNITY COST OF FREE TIME	*Marginal rate of transformation*: The amount of consumption you can gain from giving up an hour of free time, which is equal to the wage, w.	The slope of the budget constraint (the feasible frontier) which is equal to the wage.

Figure 3.19 *Your two trade-offs.*

Your optimal combination of consumption and free time is the point on the budget constraint where:

$$MRS = MRT = w$$

While considering this decision, you receive an email. A mysterious benefactor would like to give you an income of $50 a day—for life. All you have to do is provide your banking details. You realise at once that this will affect your choice of job. The new situation is shown in Figure 3.20: for each level of free time your total income—your earnings plus the mystery gift—is $50 higher than before. So the budget constraint is shifted upwards by $50—the feasible set has expanded. Your budget constraint is now:

$$c = 15(24 - t) + 50$$

Notice that the extra income of $50 does not change your opportunity cost of time: each hour of free time still reduces your consumption by $15. Your new ideal job is at B, with 19.5 hours of free time. B is the point on IC_3 where the MRS is equal to $15. With the indifference curves shown in this diagram, your response to the extra income is not simply to spend the $50; you increase consumption by less than $50, and you take some extra free time. Someone with different preferences might not choose to increase their free time: Figure 3.21 shows a case in which the MRS at each value of free time is the same on both IC_2 and the higher indifference curve IC_3. This person chooses to keep their free time the same, and consume $50 more.

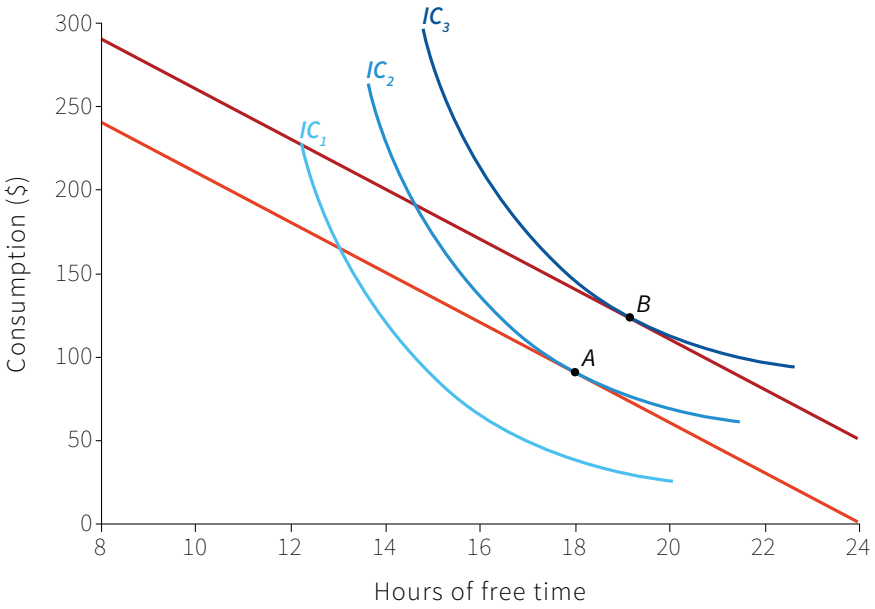


Figure 3.20 *The effect of additional income on your choice of free time and consumption.*

The effect of additional (unearned) income on the choice of free time is called the *income effect*. Your income effect, shown in Figure 3.20, is positive—that is, extra income raises your choice of free time. For the person in Figure 3.21 the income effect is zero. We assume that for most goods the income effect will be either positive or zero, but not negative: if your income increased, you would not choose to have less of something that you valued.

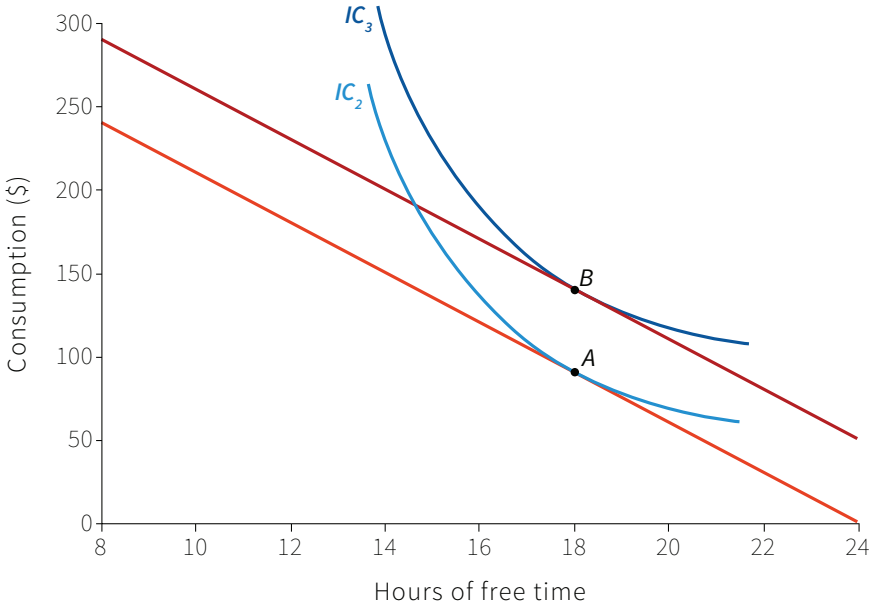


Figure 3.21 *The effect of additional income for someone whose MRS doesn't change when consumption rises.*

You suddenly realise that it might not be wise to give the mysterious stranger access to your bank account—perhaps it is a hoax. Regretfully you return to the original plan, and find a job requiring 6 hours of work per day. A year later, your fortunes improve: your employer offers a pay rise of $10 per hour, and the chance to renegotiate your hours. Now your budget constraint is:

$$c = 25(24 - t)$$

In Figure 3.22a you can see how the budget constraint changes when the wage rises. With 24 hours of free time (and no work) your consumption would be 0 whatever the wage. But for each hour of free time you give up, your consumption can now rise by $25, rather than $15. So your new budget constraint is a steeper straight line through $(24, 0)$, with a slope equal to $25. Your feasible set has expanded. And now you achieve the highest possible utility at point D, with only 17 hours of free time. So you ask your employer if you can work longer hours—a 7-hour day.

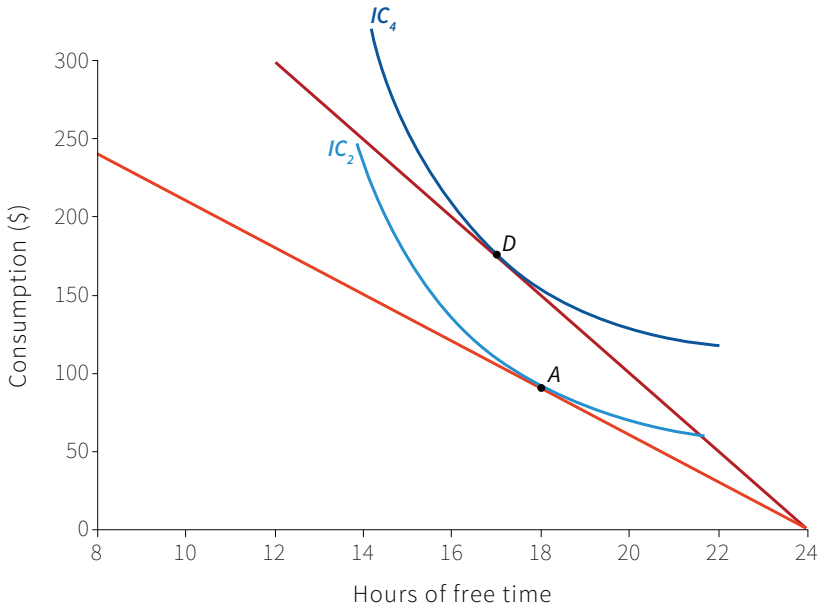


Figure 3.22a *The effect of a wage rise on your choice of free time and consumption.*

Compare the outcomes in Figures 3.21 and 3.22a. With an increase in unearned income you want to work fewer hours, while the increase in the wage in Figure 3.22a makes you decide to increase your work hours. Why does this happen? Because there are two effects of a wage increase:

- *More income for every hour worked*: For each level of free time you can have more consumption, and your MRS is higher: you are now more willing to sacrifice consumption for extra free time. This is the income effect we saw in Figure 3.21—you respond to additional income by taking more free time as well as raising your consumption.

- *The budget constraint is steeper*: But the opportunity cost of this free time is now higher. In other words, the marginal rate at which you can transform time into income (the MRT) has increased. And that means you have an incentive to work more—to decrease your free time. This is called the *substitution effect*.

The substitution effect captures the idea that when a good becomes more expensive relative to another good, you choose to substitute some of the other good for it. It is the effect a change in the opportunity cost would have on its own, for a given utility level.

We can show both of these effects in the diagram. Before the wage rise you are at A on IC_2. The higher wage enables you to reach point D on IC_4. Figure 3.22b shows how we can decompose the change from A to D into two steps corresponding to the two effects.

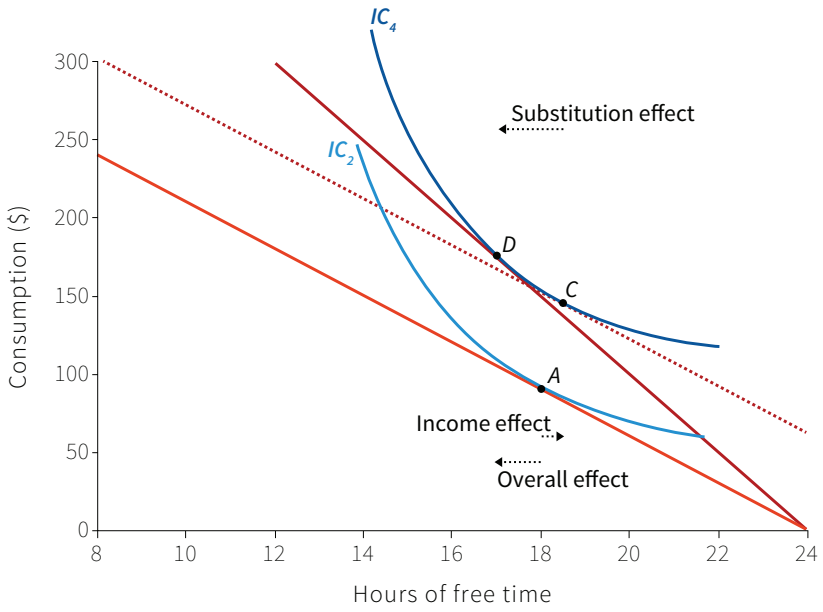


Figure 3.22b *The effect of a wage rise on your choice of free time and consumption.*

You can see in Figure 3.22b that with indifference curves of this typical shape a substitution effect will always be negative: with a higher opportunity cost of free time you choose a point on the indifference curve with a higher MRS—and that means a point with less free time (and more consumption). The overall effect of a wage rise depends on the *sum* of the income and substitution effects. In Figure 3.22b the negative substitution effect is bigger than the positive income effect, so free time falls.

INCOME AND SUBSTITUTION EFFECT

A wage rise:

- Raises your income for each level of free time, increasing the level of utility you can achieve
- Increases the opportunity cost of free time

So it has two effects on your choice of free time:

- The *income effect* (because the budget constraint shifts outwards): the effect that the additional income would have if there were no change in the opportunity cost
- The *substitution effect* (because the slope of the budget constraint—the MRT—rises): the effect of the change in the opportunity cost, given the new level of utility

If you look back at section 3.5, you will see that Angela's response to a rise in productivity was also determined by these two opposing effects: an increased incentive to work produced by the rise in the opportunity cost of free time, and (depending on her preferences) an increased desire for free time when her income rises.

We used the model of the self-sufficient farmer to see how technological change can affect working hours. Angela can respond directly to the increase in her productivity brought about by the introduction of a new technology. Employees also become more productive as a result of technological change. As we saw in Unit 2, this may lead to a rise in the wage if they have sufficient bargaining power. The model in this section suggests that, if that happens, technological progress will also bring about a change in the amount of time employees wish to spend working.

The income effect of a higher wage makes workers want more free time, while the substitution effect provides an incentive to work longer hours. If the income effect dominates the substitution effect, workers will prefer fewer hours of work.

3.8 IS THIS A GOOD MODEL?

We have looked at three different contexts in which people decide how long to spend working—a student (Alexei), a farmer (Angela), and a wage-earner (hopefully you, in the future). In each case we have modelled their preferences and feasible set, and the model tells us that their best—utility-maximising—choice is the level of working hours where the slope of the feasible frontier is equal to the slope of the indifference curve.

You may have been thinking: *this is not what people do!*

Billions of people organise their working lives without knowing anything about MRS and MRT (if they did make decisions that way, perhaps we would have to subtract the hours they would spend making calculations). And, even if they did make their choice using mathematics, most of us can't just leave work when we want. So how can this model be useful?

Remember from Unit 2 that models help us "see more by looking at less". Lack of realism is an intentional feature of this model, not a shortcoming.

Trial and error replaces calculations

Can a model that ignores how we think possibly be a good model of how we choose?

Milton Friedman, an economist, explained that when economists use models in this way they do not claim that we actually think through these calculations—we don't equate MRS to MRT—each time we make a decision. Instead we each try various choices (sometimes not even intentionally) and we tend to adopt the ones that that make us feel satisfied and not regretful about our decisions as habits, or rules of thumb.

In his book *Essays in positive economics*, he described it as similar to playing billiards (pool):

"Consider the problem of predicting the shots made by an expert billiard player. It seems not at all unreasonable that excellent predictions would be yielded by the hypothesis that the billiard player made his shots as if he knew the complicated mathematical formulas that would give the optimum directions of travel, could estimate accurately by eye the angles, etc., describing the location of the balls, could make lightning calculations from the formulas, and could then make the balls travel in the direction indicated by the formulas. Our confidence in this hypothesis is not based on the belief that billiard players, even expert ones, can or do go through the process described. It derives rather from the belief that, unless in some way or other they were capable of reaching essentially the same result, they would not in fact be expert billiard players."
—Milton Friedman, *Essays in positive economics* (1953)

Similarly, if we see a person regularly choosing to go to the library after lectures instead of going out, or not putting in much work on their farm, or asking for longer shifts after a pay rise, we do not need to suppose that this person has done the calculations we set out. If that person later regretted the choice, next time they might go out a bit more, work harder on the farm, or cut their hours back. Eventually we could speculate they might end up with a decision on work time that is close to the result of our calculations.

That is why economic theory can help to explain, and sometimes even predict, what people do—even though those people are not performing the mathematical calculations that economists make in their models.

The influence of culture and politics

A second unrealistic aspect of the model: employers typically choose working hours, not individual workers, and employers often impose a longer working day than workers prefer. As a result the hours that many people work are regulated by law, so that beyond some maximum neither the employee nor the employer can choose. In this case the government has limited the feasible set of hours and goods.

Although individual workers often have little freedom to choose their hours, it may nevertheless be the case that changes in working hours over time, and differences between countries, partly reflect the preferences of workers. If many individual workers in a democracy wish to lower their hours, they may "choose" this indirectly as voters, if not individually as workers. Or they may bargain as members of a trade union for contracts requiring employers to pay higher overtime rates for longer hours.

This explanation stresses *culture* (meaning changes in preferences or differences in preferences among countries) and *politics* (meaning differences in laws, or trade union strength and objectives). They certainly help to explain differences in working hours between countries:

- *Cultures seem to differ*. Some northern European cultures highly value their vacation times, while South Korea is famous for the long hours that employees put in.
- *Legal limits on working time differ*. In Belgium and France the normal work week is limited to 35-39 hours, while in Mexico the limit is 48 hours and in Kenya even longer.

But, even on an individual level, we may influence the hours we work. For example, employers who advertise jobs with the working hours people prefer may find they have more applicants than those offering too many (or too few) hours.

DISCUSS 3.8: ANOTHER DEFINITION OF ECONOMICS

Lionel Robbins, an economist, wrote in 1932 that:

"Economics is the science that studies human behaviour as a relationship between given ends and scarce means which have alternative uses."

1. Give an example from this unit to illustrate the way that economics studies human behaviour as a relationship between given ends and scarce means with alternative uses.
2. Are the ends of economic activity, that is, the things we desire, fixed? Think of examples from this unit—about study time and grades, or working time and consumption—to illustrate your answer.
3. The subject matter that Robbins refers to—doing the best you can in a given situation—is an essential part of economics. But is economics limited to the study of "scarce means which have alternative uses"? In answering this question, include a contrast between Robbins' definition and the one given in Unit 1 and note that Robbins wrote this passage at a time when 15% of the British workforce was unemployed.

Remember we also judge the quality of a model by whether it provides insight into something that we want to understand. In the next section, we will look at whether our model of the choice of hours of work can help us understand why working hours differ so much between countries and why, as we saw in the introduction, they have changed over time.

3.9 EXPLAINING OUR WORKING HOURS

During the year 1600 the average British worker was at work for 266 days. This statistic did not change much until the Industrial Revolution. Then, as we know from the previous unit, wages began to rise, and working time rose too: to 318 days in 1870.

Meanwhile, in the US, hours of work increased for many workers who shifted from farming to industrial jobs. In 1865 the US abolished slavery, and former slaves used their freedom to work much less. In many countries, from the late 19th century

until the middle of the 20th century working time gradually fell. Figure 3.1 at the beginning of this unit showed how annual working hours have fallen since 1870 in the Netherlands, the US and France.

The simple models we have constructed cannot tell the whole story. Remember that the *ceteris paribus* assumption can omit important details: things that we have held constant in models may vary in real life.

As we explained in the previous section, our model omitted two important explanations, which we called *culture* and *politics*. Our model provides another explanation: economics.

The economics of changes over time

Look at the two points in Figure 3.23, giving estimates of average amounts of daily free time and goods per day for employees in the US in 1900 and in 2013. The slopes of the budget constraints through points *A* and *D* are the real wage in 1900 and in 2013. This shows us the feasible sets of free time and goods that would have made these points possible. Then we consider the indifference curves of workers that would have led workers to choose the hours they did. We cannot measure indifference curves directly: we must use our best guess of what the preferences of workers would have been, given the actions that they took.

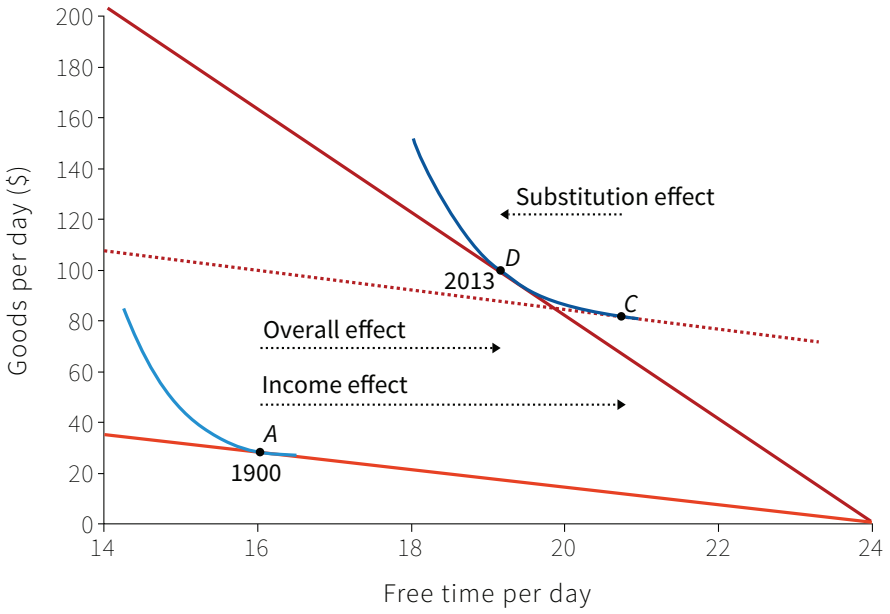


Figure 3.23 *Applying the model to history: Increased goods and free time in the US (1900-2013).*

Source: *OECD. 2015. 'Average Annual Hours Actually Worked per Worker.' Accessed June. https://stats.oecd.org/Index.aspx?DataSetCode=ANHRS. Huberman, Michael, and Chris Minns. 2007. 'The Times They Are Not Changin': Days and Hours of Work in Old and New Worlds, 1870–2000.'Explorations in Economic History 44 (4): 538–67. http://personal.lse.ac.uk/minns/Huberman_Minns_EEH_2007.pdf.*

Using the model to explain historical change

We can compare the model to the change between 1900 and 2013 in daily free time and goods per day for employees in the US. The solid lines show the feasible sets for free time and goods in 1900 and 2013, where the slope of each budget constraint is the real wage. Inferred indifference curves assume that workers chose the hours they worked. The shift from *A* to *C* is the income effect of the wage rise; on its own it would cause US workers to take more free time. The rise in the opportunity cost of free time caused US workers to choose *D* rather than *C*, with less free time.The overall effect of the wage rise depends on the sum of the income and substitution effects. In this case the income effect is bigger, so with the higher wage US workers took more free time as well as more goods.

How does our model explain how we got from point *A* to point *D*? You know from Figure 3.22 that the increase in wages would lead to both an income effect and a substitution effect. In this case, the income effect outweighs the substitution effect, so both free time and goods consumed per day go up. Figure 3.23 is thus simply an application to history of the model illustrated in Figure 3.22. Work through the steps to see the income and substitution effects.

How could reasoning in this way explain the other historical data that we have?

First, consider the period before 1870 in Britain, when both working hours and wages rose:

- *Income effect*: At the relatively low level of consumption in the period before 1870, workers' willingness to substitute free time for goods did not increase much when rising wages made higher consumption possible.
- *Substitution effect*: But they were more productive and paid more, so each hour of work brought more rewards than before in the form of goods, increasing the incentive to work long hours.
- *Substitution effect dominated*: Therefore before 1870 the negative substitution effect (free time falls) was bigger than the positive income effect (free time rises), so work hours rose.

During the 20th century we saw rising wages and falling hours. Our model accounts for this change as follows:

- *Income effect*: By the late 19th century workers had a higher level of consumption and valued free time relatively more—their marginal rate of substitution was higher—so the income effect of a wage rise was larger.
- *Substitution effect*: This was consistent with the period before 1870.

- *Income effect now dominates*: When the income effect began to outweigh the substitution effect, working time fell.

Nevertheless, the combined political, cultural and economic influences on our choices may produce some surprising trends. In our video Juliet Schor, a sociologist and economist at Boston College who has written about the paradox that many of the world's wealthiest people are working more, despite gains in technology, asks what this means for our quality of life, and for the environment.

The economics of differences between countries

Figure 3.2 showed that in countries with higher income (GDP per capita) workers tend to have more free time, but also that there are big differences in annual hours of free time between countries with similar income levels. To analyse these differences using our model we need a measure of income that corresponds more closely to earnings from employment, rather than GDP per capita. The table below shows working hours for five countries, together with the disposable income of an average employee (based on the taxes and benefits for a single person without children). From these figures we have calculated annual free time, and the average wage (by dividing annual income by annual hour worked). Finally, free time per day and daily consumption are calculated by dividing annual free time and earnings by 365.

COUNTRY	AVERAGE ANNUAL HOURS WORKED PER EMPLOYEE	AVERAGE ANNUAL DISPOSABLE INCOME (SINGLE PERSON, NO CHILDREN)	AVERAGE ANNUAL FREE TIME	WAGE (DISPOSABLE INCOME PER HOUR WORKED)	FREE TIME PER DAY	CONSUMPTION PER DAY
US	1,789	36,737	6,971	20.54	19.10	100.65
SOUTH KOREA	2,163	39,686	6,597	18.35	18.07	108.73
NETHERLANDS	1,383	40,171	7,377	29.05	20.21	110.06
TURKEY	1,855	17,118	6,905	9.23	18.92	46.90
MEXICO	2,226	11,046	6,534	4.96	17.90	30.26

Figure 3.24 *Free time and consumption per day across countries (2013).*

Source: *OECD. 2015. 'Average Annual Hours Actually Worked per Worker.' Accessed June. Net income after taxes calculated in US dollars using PPP exchange rates.*

Figure 3.25 shows how we might use this data, with the model of section 3.7, to understand the differences between the countries. From the data in Figure 3.24 we have plotted daily consumption and free time for a typical worker in each country, with the corresponding budget constraint—as before using a line through *(24,0)* with slope equal to the wage. We have no information about the preferences of workers in each country, and we don't know whether the combinations in the diagram can be interpreted as a choice made by the workers. But, if we assume that they do reflect the hours chosen by workers, we can consider what the data tells us about the preferences of workers in different countries.

From Figure 3.25, we see that average free time in Mexico and South Korea were virtually the same, although the wage was much higher in South Korea than in Mexico. South Koreans, Americans and Dutch people have about as much to spend per day, but South Koreans have three hours less free time. Could it be that South Koreans have the same preferences as Americans, so that if the wage increased in South Korea they would make the same choice? This seems unlikely: the substitution effect would lead them to consume more goods and take less free time, and it is implausible to suppose that the income effect of a wage rise would lead them to consume fewer goods. More plausible is the hypothesis that South Koreans and Americans (on average) have different preferences. Work through Figure 3.25 to see some hypothetical indifference curves that could explain the differences among countries. Notice that the indifference curves for the US and for South Korea cross. This means that South Koreans and Americans must have different preferences.

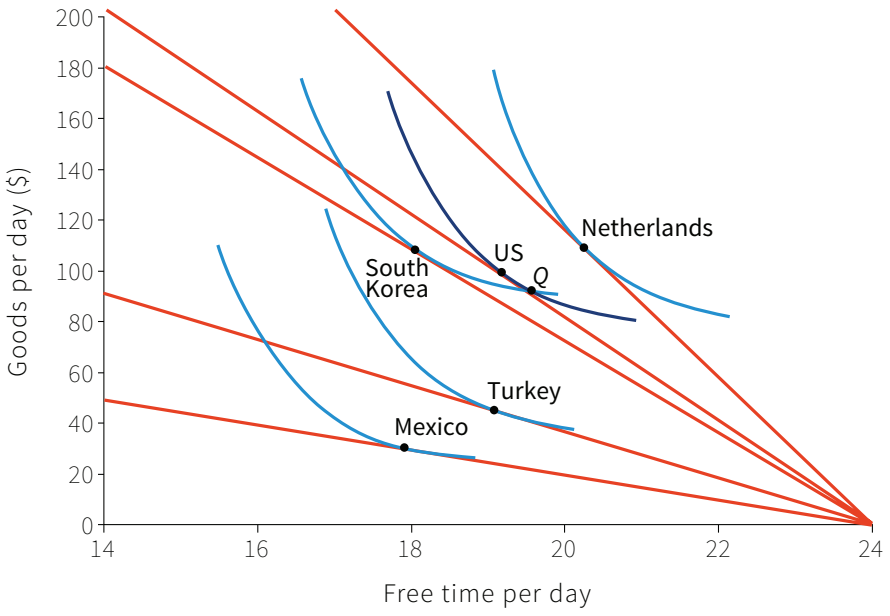


Figure 3.25 *Using the model to explain free time and consumption per day across countries (2013).*

Source: *OECD. 2015. 'Average Annual Hours Actually Worked per Worker.' Accessed June. https://stats.oecd.org/Index.aspx?DataSetCode=ANHRS. Net income after taxes calculated in US dollars using PPP exchange rates.*

Point Q in the last step of the figure is the point of intersection of the two indifference curves shown for South Korea and the US. At that point the US indifference curve is steeper than the South Korean one. This means that the average American is willing to give up more units of daily goods for an hour of free time (this is the MRS) than the average South Korean, which is consistent with the idea that South Koreans work exceptionally hard. This shows that it may be important to take account of differences in preferences among countries, or among individuals.

The possibility that our preferences change over time is equally important. Figure 3.1 is an example. Look carefully and you can see that in the last part of the 20th century hours of work rose in the US, even though wages hardly rose. Hours of work also increased in Sweden during this period.

Why? Perhaps Swedes and Americans came to value consumption more over these years. If this occurred, their preferences changed so that their MRS fell (they became more like today's South Korean workers). This may have occurred because in both the US and Sweden the share of income gained by the very rich increased considerably, and the lavish consumption habits of the rich set a higher standard for everyone else. According to this explanation Swedes and Americans were "keeping up with the Joneses" and the Joneses got richer, leading everyone else to change their preferences.

DISCUSS 3.9: PREFERENCES AND CULTURE

Suppose that the points plotted in Figure 3.25 reflect the choices of free time and consumption made by workers in these five countries according to our model.

1. Is it possible that people in Turkey and the US have the same preferences? If so, how will a wage rise in Turkey affect consumption and free time? What does this imply about the income and substitution effects?
2. Suppose that people in Turkey and South Korea have the same preferences. In that case, what can you say about the income and substitution effects of a wage increase?
3. If wages in South Korea increased, would you expect consumption there to be higher or lower than in the Netherlands? Why?

DISCUSS 3.10: WORKING HOURS ACROSS COUNTRIES AND TIME

To see what has happened to work hours in many countries during the 20th century, look at this data.

1. How would you describe what happened?
2. How are the countries in Panel A of the figure different from those in Panel B?
3. What possible explanations can you suggest for why the decline in work hours was greater in some countries than in others?
4. Why do you think that the decline in work hours is faster in most countries in the first half of the century?
5. In recent years, is there any country in which working hours have increased? Why do you think this happened?

3.10 CONCLUSION

Over the past century, hours of work have fallen, but not by as much as John Maynard Keynes, a British economist, predicted. In 1930 he published *Economic possibilities for our grandchildren*, in which he suggested that in the 100 years that would follow, technological improvement would make us, on average, about eight times better off. What he called "the economic problem, the struggle for subsistence" would be solved; we would not have to work more than, say, 15 hours per week to satisfy our economic needs. The question he raised was: how would we cope with all of the additional leisure time?

Keynes' prediction for the rate of technological progress in countries such as the UK and the US has been approximately right, but it seems very unlikely that working hours will have fallen to 15 hours per week by 2030.

Nevertheless the high-income economies will continue to experience a major transformation: the declining role of work in the course of our lifetimes. We go to work at a later age, stop working at an earlier age of our longer lives, and spend fewer hours at work during our working years. Robert Fogel, an economic historian, estimated the total working time, including travel to and from work, and housework in the past. He made projections for the year 2040, defining what he called *discretionary time* as 24 hours a day minus the amount we all need for what a kind

of biological maintenance (sleeping, eating and personal hygiene). Fogel counted leisure time as discretionary time minus working time.

In 1880 he estimated that lifetime leisure time was just a quarter of lifetime work hours. In 1995 leisure time exceeded working time over a person's entire life. He predicted that lifetime leisure would be three times lifetime work hours by the year 2040. His estimates are in Figure 3.26.

We do not yet know if Fogel has overstated the future decline in working time, as Keynes once did. But he certainly is right that one of the great changes brought about by the technological revolution is the vastly reduced role of work in the life of an average person. We return to this question in Unit 20.

CONCEPTS INTRODUCED IN UNIT 3

Before you move on, review these definitions:

- Constrained choice problem
- Scarcity
- Opportunity cost
- Marginal product
- Indifference curve
- Marginal rate of substitution (MRS)
- Marginal rate of transformation (MRT)
- Feasible set
- Budget constraint
- Income effect
- Substitution effect

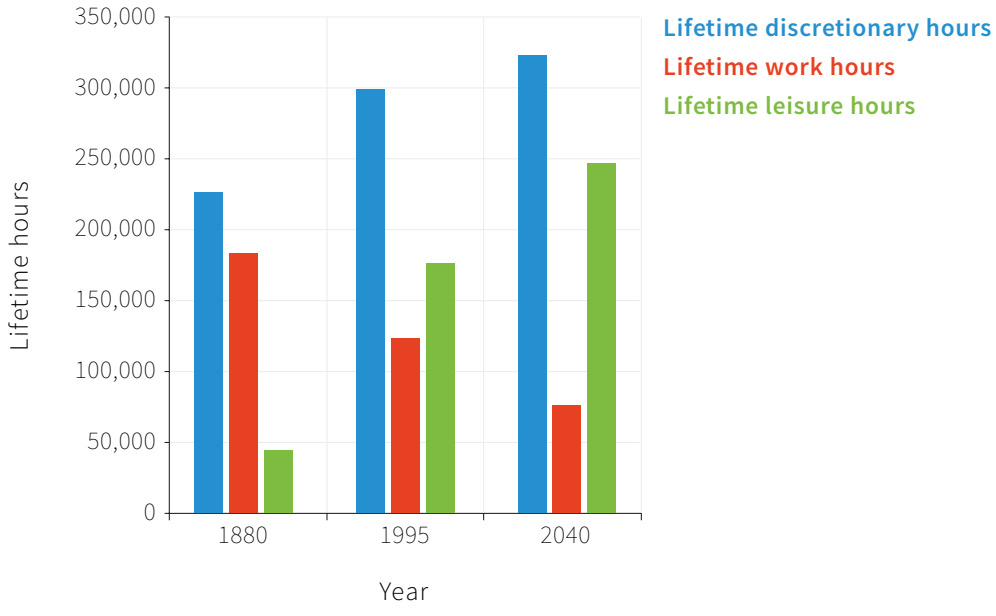


Figure 3.26 *Estimated lifetime hours of work and leisure (1880, 1995, 2040).*

Source: *Fogel, Robert W. 2000. The Fourth Great Awakening and the Future of Egalitarianism. Chicago, IL and London: University Of Chicago Press.*

DISCUSS 3.11: SCARCITY AND CHOICE

1. Do our models of scarcity and choice provide a plausible explanation for the trends in working hours during the 20th century?
2. What other factors, not included in the model, might be important in explaining what has happened?
3. Why do you think working hours since 1930 have not changed as Keynes expected? Have people's preferences changed? The model focuses on the number of hours workers would choose; do you think that many employees are now working longer than they would like?
4. In his essay, Keynes said that people have two types of economic needs or wants: *absolute needs* that we feel whatever the situation of our fellow humans, and *relative needs*—which he called "the desire for superiority". The phrase "keeping up with the Joneses" captures a similar idea that our preferences could be affected by observing the consumption of others. Could relative needs help to explain why Keynes was so wrong about working hours?

Key points in Unit 3

Decision-making under scarcity

The model of decision-making under scarcity can be applied to problems for which the ways in which we satisfy our objectives are limited by the means at our disposal.

Changes in work hours

This model can be used to understand some of the reasons for changes in hours of work over the last century, and differences in work hours across countries.

Indifference curves and the feasible set

One's preferences are described by a set of indifference curves, while one's choices are limited by the boundary of the feasible set.

MRS and MRT

In deciding how many hours to work, a person has to balance a trade-off based on the relative desirability of consumption and free time (represented by the MRS, the slope of the indifference curve) against a trade-off based on the feasible set (represented by the MRT, the slope of the feasible frontier).

Income and substitution effects

The effect of a change in the feasible set on an individual's choices will generally include both an income effect and a substitution effect.

Improvements in technology and wages

An improvement in technology or an increase in wages is likely to alter the marginal rate of transformation between goods and free time, raising the opportunity cost of free time.

A higher MRT gives workers an incentive to work longer hours (the substitution effect). But higher income may increase their desire for free time (the income effect). The overall change in hours of work depends on which of these effects is bigger.

Limitations of the model

Like all models, the model of work hours excludes potentially important factors, such as differences in preferences and legislation as influences on work hours.

3.11 EINSTEIN

How much free time?

The example in the introduction asks you to imagine that you earn $15 an hour for a 40-hour working week: your earnings are $600 per week. There are 24 hours in a day and 168 hours in a week so, after 40 hours of work, you are left with 128 hours of free time.

Your hourly wage rises to $90 and your prospective employer lets you choose how many hours you work each week.

If this were your choice, you would enjoy an additional 33 hours and 20 minutes (about 26%) more free time than previously. How do we work this out?

Suppose you are happy earning just $600 per week, and so convert any hourly pay increase into more free time. At an hourly wage rate of $90 per hour you can earn $600 by working:

$$\frac{600}{90} = 6.67 \text{ hours}$$

That is, 6 hours and 40 minutes a week. Hence your consumption of free time in a week will rise by 33 hours and 20 minutes from 7 x 24 - 40 = 128 hours to 7 x 24 - 6.67 = 161.33 hours. That is, by:

$$\frac{161.33 - 128}{128} \times 100 = 26\%$$

3.12 READ MORE

Bibliography

1. Burgoon, Brian, and Phineas Baxandall. 2004. 'Three Worlds of Working Time: The Partisan and Welfare Politics of Work Hours in Industrialized Countries.' *Politics & Society* 32 (4): 439–73.
2. Fogel, Robert W. 2000. *The Fourth Great Awakening and the Future of Egalitarianism*. Chicago, IL and London: University Of Chicago Press.
3. Friedman, Milton. (1953) 1966. *Essays in Positive Economics*. Chicago, IL: University of Chicago Press.

4. Harford, Tim. 2015. 'The Rewards for Working Hard Are Too Big for Keynes's Vision.' *The Undercover Economist*.
5. Huberman, Michael, and Chris Minns. 2007. 'The Times They Are Not Changin': Days and Hours of Work in Old and New Worlds, 1870–2000.' *Explorations in Economic History* 44 (4): 538–67.
6. Keynes, John Maynard. (1930) 1963. 'Economic Possibilities for Our Grandchildren.' In *Essays in Persuasion*. New York, NY: W. W. Norton & Co.
7. Maddison Project. 2013. '2013 Edition'.
8. OECD. 2015. 'Average Annual Hours Actually Worked per Worker.' Accessed June.
9. OECD. 2015. 'Level of GDP per Capita and Productivity.' Accessed June.
10. Plant, Ashby E., Anders K. Ericsson, Len Hill, and Kia Asberg. 2005. 'Why Study Time Does Not Predict Grade Point Average across College Students: Implications of Deliberate Practice for Academic Performance.' *Contemporary Educational Psychology* 30 (1): 96–116.
11. Robbins, Lionel. 1932. 'An Essay on the Nature and Significance of Economic Science.' London: Macmillan and Co.
12. Schor, Juliet B. 1991. *The Overworked American: The Unexpected Decline of Leisure*. New York, NY: Basic Books.
13. Veblen, Theodore. (1899) 2009. *The Theory of the Leisure Class*. Oxford: Oxford University Press.

Beta September 2015 version

4

SOCIAL INTERACTIONS

Les Joueurs de Carte, Paul Cézanne, 1892-95, Courtauld Institute of Art

A COMBINATION OF SELF-INTEREST, A REGARD FOR THE WELLBEING OF OTHERS, AND APPROPRIATE INSTITUTIONS CAN YIELD DESIRABLE SOCIAL OUTCOMES WHEN PEOPLE INTERACT

- Game theory is a way of understanding how people interact based on the constraints that limit their actions, their motives and their beliefs about what others will do
- Experiments and other evidence show that self-interest, a concern for others and a preference for fairness are all important motives explaining how people interact
- In most interactions there is some conflict of interest between people, and also some opportunity for mutual gain
- The pursuit of self-interest can lead either to results that are considered good by all participants, or sometimes to outcomes that none of those concerned would prefer
- Self-interest can be harnessed for the general good in markets by governments limiting the actions that people are free to take, and by one's peers imposing punishments on actions that lead to bad outcomes
- A concern for others and for fairness allows us to internalise the effects of our actions on others, and so can contribute to good social outcomes

Institute for New Economic Thinking
UCL SciencesPo Azim Premji University FRIENDS PROVIDENT Foundation

"The scientific evidence is now overwhelming: climate change presents very serious global risks, and it demands an urgent global response". This is the blunt beginning of the executive summary of a document called the *Stern Review*, published in 2006. The British chancellor of the exchequer (finance minister) commissioned a group of economists, led by former World Bank chief economist Sir Nicholas (now Lord) Stern, to assess the evidence for climate change, and to try to understand its economic implications. The *Stern Review* predicts that the benefits of early action will outweigh the costs.

The *Fifth Assessment Report* by the Intergovernmental Panel on Climate Change (IPCC) agrees. This early action would mean a significant cut in greenhouse gas emissions, requiring a reduction in the quantity of energy-intensive goods we consume, a switch to different energy technologies, and an improvement in the efficiency of current technologies.

But none of this will happen if we pursue what Stern referred to as "business as usual". (Remember this phrase: we will use it again at the end of this unit.) This is a scenario in which people, governments and businesses are free to pursue their own pleasures, politics and profits without taking more adequate account of the effect of their actions on others, including future generations.

National governments conflict on the policies that should be adopted. Many European nations are pressing for strict global controls on carbon emissions, while India and China, whose economic catch-up with Europe has been dependent on coal-burning technologies, resist these measures.

The problem of climate change is far from unique. It is an example of what is called a *social dilemma*. Social dilemmas—like climate change—occur when people do not take adequate account of the effects of their decisions on others, whether these are positive or negative.

Social dilemmas occur frequently in our lives. Traffic jams happen when our choice of a way to get around—for example driving alone to work rather than car-pooling—do not take account of the contribution to congestion that we make. We overuse antibiotics for minor illnesses: the sick person who takes antibiotics recovers more quickly, but overuse creates antibiotic-resistant bacteria that have a much more harmful effect on many others.

SOCIAL DILEMMA

A situation in which:

- Actions, taken independently by individuals
- ... in pursuit of their own private objectives
- Result in an outcome that is inferior to some other feasible outcome that could have resulted
- ... had people acted together, rather than as individuals

In 1968 Garrett Hardin, a biologist, published an article about social dilemmas in the journal *Science*, called *The Tragedy of the Commons*. He noted that resources that are not owned by anyone, such as the earth's atmosphere or fish stocks, are easily overexploited unless we control access in some way. Fishermen as a group would be better off not catching as much tuna, and consumers as a whole would be better off not eating it. Humanity would be better off by emitting less pollutants, but if you, as an individual, decide to cut your consumption, your carbon footprint, or the number of tuna you catch, your sacrifice will hardly make a dent in the global problem.

Examples of Hardin's tragedies, and other social dilemmas, are all around us: if you live with roommates, or in a family, you know just how difficult it is to keep a clean kitchen or bathroom. When one person cleans everyone benefits; but it is hard work. Whoever cleans up bears this cost; the others are sometimes called *free riders*. If as a student you have ever done a group assignment, you understand that the cost of effort (to gather evidence, or write up the results, or think about the problem) is individual, yet the benefits (a better grade, a higher class standing, or simply the admiration of classmates) go to the whole group.

There is nothing new about social dilemmas; we have been facing them since prehistory. Sometimes we solve them, but sometimes not (or not yet), as in the case of climate change.

More than 2,500 years ago, the Greek storyteller Aesop wrote about a social dilemma in his fable *Belling the Cat*, in which a group of mice needs one of its members to place a bell around a cat's neck. Once the bell is on, the cat cannot catch and eat the other mice; but the outcome may not be so good for the mouse that takes the job. There are countless examples during wars, or in natural catastrophes, in which individuals sacrifice their lives for others who are not family members, and may even be total strangers. These actions are termed *altruistic*.

Altruistic self-sacrifice is not the most important way societies resolve social dilemmas and reduce free riding. Sometimes the problems can be resolved by government policies; for example governments have successfully imposed quotas to prevent the over-exploitation of stocks of cod in the North Atlantic. In the UK, the amount of waste that is dumped in landfill, rather than being recycled, has been dramatically reduced by a landfill tax.

Local communities also create institutions to regulate behaviour. Irrigation communities need people to work to maintain the canals that benefit the whole community. Individuals also need to use scarce water sparingly so that other crops will flourish, although this will lead to smaller crops for the individual. In Valencia, Spain, communities of farmers have used a set of customary rules for centuries to regulate communal tasks and to avoid using too much water. Since the middle ages they have had an arbitration court called the *Tribunal de las Aguas* (Water Court) that resolves conflicts between farmers about the application of the rules. The ruling of the Tribunal is not legally enforceable. Its power comes only from the respect of the community, yet its decisions are almost universally followed.

Even present-day global environmental problems have sometimes been tackled effectively. The *Montreal Protocol* has been remarkably successful. It was created to phase out and eventually ban the chlorofluorocarbons (CFCs) that threatened to destroy the ozone layer that protects us against harmful ultraviolet radiation.

Sometimes self-interest, when properly channelled, can be as much part of the solution to social dilemmas as part of the problem. For example, in Unit 1 and Unit 2 we saw how economic self-interest can contribute to social wellbeing when entrepreneurs have incentives to develop new technologies, and to copy those who innovate. In a well-functioning capitalist economy (check Figure 1.11 in Unit 1 to remind yourself of the conditions for this) there are many markets in which total strangers buy and sell goods, each being motivated by their own benefits, but also creating a benefit for the people on the other side of the transaction.

With the right institutions, self-interest can be channelled so that the result is, for the most part, mutually beneficial. This is one of the most important principles of economics, but it was a startling idea when Adam Smith described it in 1759. The phrase he used to describe the social benefit of individual self-interest is still common today: he called it "the invisible hand".

DISCUSS 4.1: SOCIAL DILEMMAS

Using the news headlines from last week:

1. Identify two social dilemmas that have been reported (try to use examples not discussed above).
2. For each, specify how it satisfies the definition of a *social dilemma*.

4.1 SOCIAL INTERACTIONS: GAME THEORY

On which side of the road should you drive? If you live in Japan, the UK or Indonesia, you drive on the left. If you live in South Korea, France or the US, you drive on the right. If you grew up in Sweden, you drove on the left until 5pm on 3 September 1967, when the law changed, and on the right afterwards. The government sets a rule, and we follow it.

But suppose we just left the choice to drivers to pursue their self-interest and to select one side of the road or the other. If everyone else was already driving on the right, self-interest (avoiding a collision) would be sufficient to motivate a driver to drive on the right as well. Concern for other drivers, or a desire to obey the law, would not be necessary.

Devising policies to promote people's wellbeing requires an understanding of the difference between situations in which self-interest can promote general wellbeing, and cases in which it leads to undesirable results. To do this we will introduce *game theory*, a way of modelling how people interact.

In Unit 3 we saw how a student deciding how much to study, or a farmer how hard to work, faces a set of feasible options. This person then makes decisions to obtain the best possible outcome. In both cases the feasible outcomes were determined by a production function specifying a relationship between the amount of work done, and the result.

But in the models we have studied so far, the outcome *did not depend on what anyone else did*. The student and the farmer were not engaged in a *social interaction*.

Social and strategic interactions

In this unit we consider these social interactions, meaning situations in which there are many people, and the actions taken by each person affects that person's outcome, and other people's outcomes too. For example, one person's choice of how to heat his or her home will affect another's experience of global climate change.

We use four terms:

- When people are engaged in a social interaction and are aware of the ways that their actions affect others, and vice versa, we call this a *strategic interaction*.
- A *strategy* is defined as an action (or a course of action) that a person may take when that person is aware of the mutual dependence of the results for herself and for others. The outcomes depend not only on that person's actions, but also on the actions of others.
- Models of strategic interactions are described as *games*.
- *Game theory* is a set of models of strategic interactions. It is widely used in economics and elsewhere in the social sciences.

To see how game theory can clarify strategic interactions, imagine two farmers, who we will call Anil and Bala. They face a problem: should they grow rice or cassava? Both could grow either crop, but we will assume it is never worthwhile for either of them to grow a bit of each.

Anil's land is better suited for growing cassava, while Bala's is better suited for rice. The two farmers have to determine what is called the *division of labour*, that is, who will do what to produce the crop. They do this *independently*, which means they do not agree jointly on a course of action.

(This condition of independence may seem odd in the case of just two farmers, but later we apply the same logic to situations like climate change in which hundreds or even millions of people interact, most of them total strangers to one another. So assuming that Anil and Bala do not come to some common agreement is useful for us.)

They both sell whatever crop they produce in a nearby village market. On market day, if they bring less rice to the market, the price will be higher. The same goes for cassava. Figure 4.1 describes their interaction, which is what we call a game. Let's explain what Figure 4.1 means, because you will be seeing this a lot.

Anil's choices are the rows of the table; Bala's are the columns. We call Anil the *row player* and Bala the *column player*.

When an interaction is represented in a table like Figure 4.1, it is important to think of each entry as the result of a *hypothetical situation*. For example, read the upper left cell as:

"Suppose (for whatever reason) Anil planted rice and Bala planted rice too. What would we see?"

Figure 4.1 lists all of the possible things they might do. In this case, there are four possible hypothetical situations. Then we ask, "Why would they do that?"

		Bala	
		RICE	CASSAVA
Anil	RICE	Both produce rice: There is a glut of rice on the market (it will sell for a low price) There is a shortage of cassava Anil not producing cassava, which he is better able to produce	No market glut: High prices for both crops Both farmers producing the crop for which they are less suited
	CASSAVA	No market glut: High prices for both crops Both farmers producing the crop for which they are better suited	Both produce cassava: There is a glut of cassava (low price) There is a shortage of rice Bala not producing rice, which he is better able to produce

Figure 4.1 *Social interactions in the invisible hand game.*

To simplify the model, we have assumed that:

- There are no other people involved or affected in any way.
- The selection of which crop to grow is the only decision that Anil and Bala need to make.
- At this point we assume that Anil and Bala will interact just once (this is called a *one-shot game*).

Because they decide *independently* which crop to grow, the four possible situations are described by the cells in Figure 4.2a. They also decide *simultaneously* so, when they make a decision, they don't know what the other person has decided to do.

GAME

A description of a social interaction, which specifies:

- *The players*: Who is interacting with whom
- *The feasible strategies*: Which actions are open to the players
- *The information*: Who knows what, when making the decision
- *The payoffs*: What the outcomes will be for each of the possible combinations of actions

In Figure 4.2b we show the outcomes for the two players, which are called *payoffs*, and shown in what is called a payoff matrix. A *matrix* is just any rectangular (in this case square) array of numbers. As in Figure 4.1, each cell indicates the combinations of the actions taken by Anil and Bala; but here the numbers in the cell are the income that the two will receive if the hypothetical actions in the column and row were taken. To understand the entries in the payoff matrix, remember that Anil's land is better suited to cassava and Bala's to rice. The first number is the reward received by the row player (the row player's name begins with A as a reminder that his payoff is first). The second number is the column player's payoff.

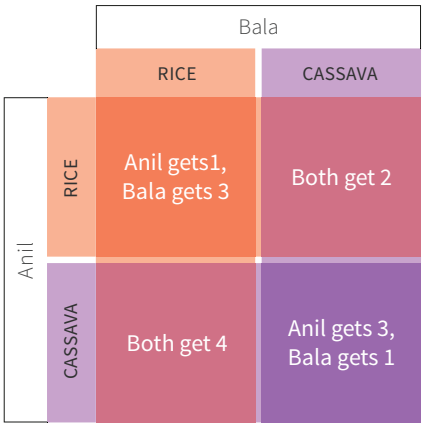


Figure 4.2a *The four possible situations in the invisible hand game.*

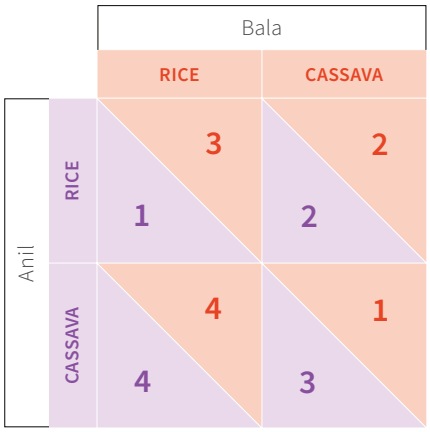


Figure 4.2b *The equivalent payoff matrix in the invisible hand game.*

The payoff matrix for this game illustrates two problems facing the two players:

- Because the market price falls when it is flooded with one crop, they can do better if they specialise, rather than both produce the same good.
- When they produce different goods they would do better to specialise in the crop for which their land is most suited.

4.2 THE DIVISION OF LABOUR AND THE INVISIBLE HAND

Game theory describes social interactions, but it also sometimes provides predictions about outcomes. The outcome of a game is just a description of the actions taken by each player, which along with Figure 4.2 gives the payoffs of each player. To predict an outcome we need another term: *best response*. This is the strategy that will yield the highest payoff, given the strategy the other person selects.

Think about best responses in the invisible hand game. Suppose you are Anil, and you consider the hypothetical case in which Bala has chosen to grow rice. What response yields you the higher payoff? You would grow cassava (Anil would then get 4; if he also grew rice, he would get a payoff of only 1).

What would Anil's best response be, given Bala's hypothetical choice of growing cassava? Anil would again grow cassava (he would get 3, rather than the 2 he would get were he to grow rice).

So, whatever Bala does, Anil's best response is to grow cassava. You can check your understanding of this game by explaining to yourself why Bala's best response is to grow rice, no matter which strategy Anil uses.

In this case, a *dominant strategy* is *Cassava* for Anil and *Rice* for Bala. The term refers to a strategy that is a best response to each of the other players' possible strategies. Because both have a dominant strategy we have a simple prediction about what each will do: play their dominant strategy.

An outcome of a game in which each player plays his or her dominant strategy is termed a *dominant strategy equilibrium*.

As in Unit 2, equilibrium means that something of interest does not change. In this case the strategies adopted by the players do not change: if they are both playing their dominant strategy, neither would have an incentive to switch to any other strategy. When there is a dominant strategy equilibrium, we can predict what the players will do.

This is how to check whether there is a dominant strategy equilibrium in a game, using a payoff matrix:

1. *Work out the best response of the row player*. If the row player's best response is the same for each of the column player's actions, then this is dominant strategy for the row player.
2. *Work out the best response of the column player*. If the column player's best response is the same for each of the row player's actions, then this is dominant strategy for the column player.
3. *If there are two dominant strategies*, the outcome of a two-player game is a dominant strategy equilibrium.

Because both Anil and Bala have a dominant strategy, neither cares what the other decides. This is similar to the models in Unit 3 in which Alexei's choice of hours of study, or Angela's working hours, did not depend on what others did. But here, while the choice of crop does not depend on what the others do, the payoffs they get from their strategy depend on the other player's decision. For example, if Anil is playing his dominant strategy (*Cassava*) he is better off if Bala also plays his dominant strategy (*Rice*), than if Bala plants cassava as well.

In the dominant strategy equilibrium Anil and Bala have specialised, rather than produce the same good, and they have specialised in the production of the good for which their land is better suited. In this case simply pursuing their self-interest—choosing the strategy for which they got the highest payoff—resulted in an outcome that was:

- The best of the four possible outcomes for each player
- The strategy that yielded the largest total payoffs for the two farmers combined

In this case the dominant strategy equilibrium is the outcome that each would have chosen if they had a way of coordinating their decisions, implementing the two strategies jointly. This is why we have called this the "invisible hand" game—although they *independently* pursued their self-interest, they were guided, as if by an invisible hand, to an outcome that was in their joint best interests.

Real economic problems are never this simple, but the basic logic is the same. The pursuit of self-interest without regard for others is sometimes considered to be morally bad, but the study of economics has identified cases in which it can lead to outcomes that are socially desirable. There are also cases, however, in which the pursuit of self-interest leads to results that are not in the self-interest of any of the players. The prisoners' dilemma game, which we study next, describes one of these situations.

WHEN ECONOMISTS DISAGREE

HOMO ECONOMICUS IN QUESTION: ARE PEOPLE ENTIRELY SELFISH?

For centuries, economists and just about everyone else have debated whether people are entirely self-interested or are sometimes happy to help others even when it costs them something to do so. *Homo economicus* (economic man) is the nickname given to the selfish and calculating character that you find in economics textbooks. Have economists been right to imagine *homo economicus* as the only actor on the economic stage?

In the same book in which he first used the phrase "invisible hand", Adam's Smith also made it clear that he though we were not *homo economicus*:

"How selfish soever man may be supposed, there are evidently some principles in his nature that interest him in the fortunes of others, and render their happiness necessary to him, though he derives nothing from it except the pleasure of seeing it."
Adam Smith, *The Theory of Moral Sentiments* (1759)

But most economists since Smith have disagreed. In 1881, F. Y. Edgeworth, a founder of modern economics, made this perfectly clear in his book *Mathematical Psychics*: "The first principle of economics is that every agent is actuated only by self-interest."

Yet everyone has experienced, and sometimes even performed, great acts of kindness or bravery on behalf of others in situations in which there was little chance of a reward. The question for economists is: should the unselfishness evident in these acts be part of how we reason about behaviour?

Some say "no": many seemingly generous acts are better understood as attempts to gain a favourable reputation among others that will benefit the actor in the future. Maybe helping others and observing social norms is just self-interest with a long time horizon. This is what the essayist H. L. Mencken thought: "conscience is the inner voice which warns that somebody may be looking."

Since the 1990s, in an attempt to resolve the debate on empirical grounds, economists have performed hundreds of experiments all over the world in which the behaviour of individuals (students, farmers, whale hunters, warehouse workers and CEOs) can be observed as they make real choices about sharing, using economic games.

In these experiments, we almost always see some self-interested behaviour. But we also observe *altruism, reciprocity, aversion to inequality*, and other preferences that are different from self-interest. In many experiments *homo economicus* is in a minority. This is true even when the amounts being shared (or kept for oneself) amount to many days' wages.

Is the debate resolved? Many economists think so and now consider, in addition to *homo economicus*, people who are sometimes altruistic, sometimes inequality averse and sometimes reciprocal. They point out that the assumption of self-interest is appropriate for many economic settings, like shopping, and the way that firms use technology to maximise profits. But it's not as appropriate in other settings, such as how we pay taxes, or why we work hard for our employer.

4.3 THE PRISONERS' DILEMMA

Imagine that Anil and Bala are now facing a different problem. Each is deciding how to deal with pest insects that destroy the crops they cultivate in their adjacent fields. Each has two feasible strategies:

- The first is to use an inexpensive chemical called *Terminator*. It kills every insect for miles around. *Terminator* also leaks into the water supply that they both use.
- The second is to use integrated pest control (IPC) instead of a chemical. A farmer using *IPC* introduces beneficial insects to the farm. The beneficial insects eat the pest insects.

If just one of them chooses *Terminator*, the damage is quite limited. If they both choose it, water contamination becomes a serious problem, and they need to buy a costly filtering system. Figure 4.3a and Figure 4.3b describe their interaction.

Both Anil and Bala are aware of these outcomes. As a result they know the amount of money they will make at harvest time, net of the costs of their pest control strategy and the installation of water filtration if that becomes necessary, will depend not only on what choice they make, *but also on the other's choice*. This is a strategic interaction.

How will they play the game? To figure this out we again use the idea of a best response.

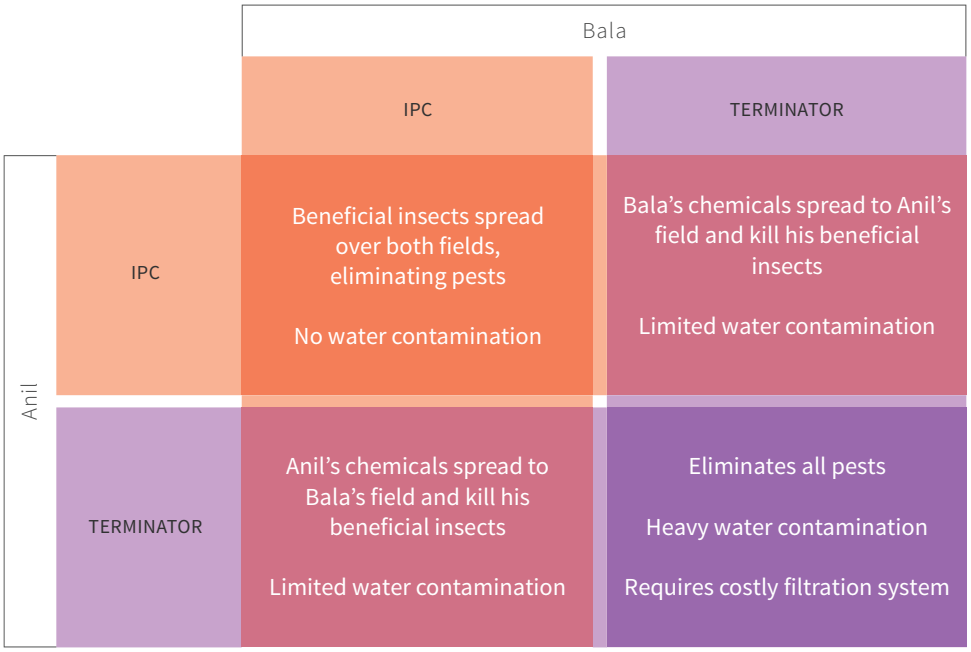


Figure 4.3a *Social interactions in the pest control game.*

- *What is Anil's best response to Bala's hypothetical use of IPC*? It is to use *Terminator* (he gets 4 in this case, rather than 3 were he to choose IPC).
- *What would Anil's best response be to Bala's hypothetical choice of using Terminator*? Anil would again use *Terminator* because, if Bala uses it, Anil could not use *IPC* on his own field: Bala's chemicals would kill off Anil's beneficial insects.

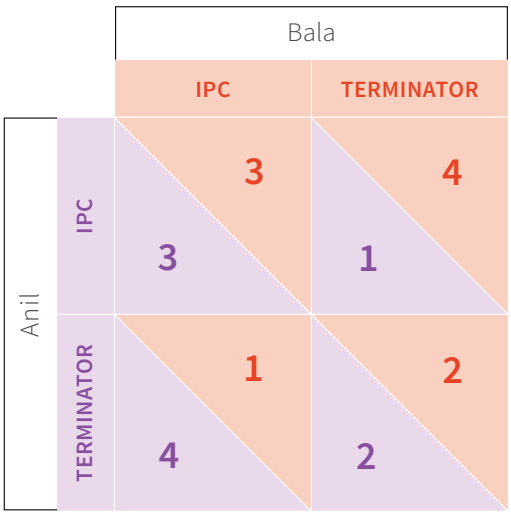


Figure 4.3b *Payoffs in the pest control game.*

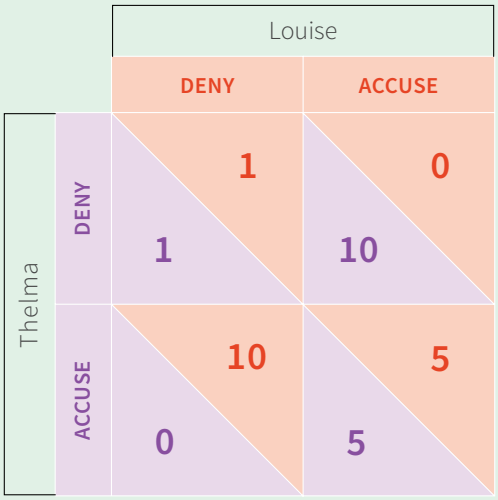
So whatever Bala does, the best response for Anil is to use *Terminator*. (You can check your understanding of this game by explaining to yourself that the same is true of Bala: his best response is to use *Terminator*, no matter which strategy Anil uses.)

This means that *Terminator* is the dominant strategy and, because *Terminator* is the dominant strategy for both, both using the insecticide is the dominant strategy equilibrium. The prediction of the game is that both will use it.

Both would be better off had they both used *IPC*. So the predicted outcome is not the best that is feasible. The pest control game is a particular example of a game called the *prisoners' dilemma*.

THE PRISONER'S DILEMMA

The name of this game comes from a story about two prisoners (we call them Thelma and Louise) whose strategies are either to *Accuse* (implicate) the other in a crime that the prisoners may have committed together, or *Deny* that the other prisoner was involved.



If both Thelma and Louise deny it, they are freed after a few days of questioning.

Accusing the other person, while the other person denies, leads the accuser to be freed immediately (a sentence of zero years), whereas the other person gets a long jail sentence (10 years).

Finally, when both Thelma and Louise choose Accuse (meaning each implicates the other), they both get a jail sentence. This sentence is reduced from 10 to five years, because of their cooperation with the police. The payoffs of the game are shown in Figure 4.4.

(The payoffs are written in terms of years of prison—so a high number is worse for Louise or Thelma's wellbeing.)

In a prisoners' dilemma, both players have a dominant strategy—in the example, *Accuse*—which, when played by both results in an outcome that is worse for both than had they both adopted a different strategy (in the example, *Deny*).

Our story about Thelma and Louise is hypothetical, but this game applies to many real problems. For example, watch this clip from the TV quiz show *Golden Balls*, and you will see how one ordinary person ingeniously resolves the prisoners' dilemma. In economic examples, the mutually beneficial strategy—*Deny*—is generally termed *Cooperate*, while the dominant strategy—*Accuse*—is called *Defect*. As in the case of Anil and Bala, Cooperate does not mean the prisoners get together and discuss what to do. The rules of the game are always that each player decides independently on a strategy.

The prisoners' dilemma does not show that self-interest necessarily leads to outcomes that nobody would endorse, any more than the invisible hand game suggests that self-interest is always best. We will see that both the prisoners' dilemma and the invisible hand game help us understand more precisely how markets can harness self-interest to improve the workings of the economy—and how sometimes they fail to do this.

PRISONER'S DILEMMA

A game in which, in the dominant strategy equilibrium:

- Payoffs are lower for each player
- Payoffs are lower in total than if neither player played the dominant strategy

Three aspects of the interaction between Anil and Bala caused us to predict an unfortunate outcome in their prisoners' dilemma game:

- Anil and Bala did not place any value on the payoffs of the other person, and so did not internalise the costs that their actions inflicted on the other.
- There was no way that Anil, Bala or anyone else could make the farmer who used the insecticide pay for the harm that it caused.
- Anil and Bala were not able to make an agreement about what each would do. Had they been able to do so, they could have simply agreed to use *IPC*, or banned the use of *Terminator*.

If we can overcome one or more of these problems, the outcome preferred by both of them would sometimes result. So, in the rest of this unit, we will explain the ways to do this.

DISCUSS 4.2: POLITICAL ADVERTISING

Many people consider political advertising (campaign advertisements) to be a classic example of a prisoner's dilemma.

1. Using examples from a recent political campaign with which you are familiar, explain whether this is the case.
2. Write down an example payoff matrix for this case.

4.4 SOCIAL PREFERENCES: ALTRUISM

When students play one-shot prisoners' dilemma games in classroom or laboratory experiments—sometimes for substantial sums of real money—it is not unusual to observe half or more of the participants playing the *Cooperate* rather than *Defect* strategy, despite mutual defection being the dominant strategy for players who care only about their own monetary payoffs. One interpretation of these results is that players are *altruistic*.

If Anil, for example had cared sufficiently about the harm that he would inflict on Bala by using *Terminator* when Bala was using *IPC*, then *IPC* would have been Anil's best response to Bala's IPC. And if Bala had felt the same way, then *IPC* would have been a mutual best response, and the two would no longer have been in a prisoners' dilemma.

A person who is willing to bear a cost in order to help another person is said to have *altruistic preferences*. In the example just given, Anil was willing forego gaining 1 payoff unit because that would have imposed a loss of 2 on Bala. His opportunity cost of choosing *IPC* when Bala had chosen *IPC* was 1, and it conferred a benefit of 2 on Bala, meaning that he had acted altruistically.

In Unit 3 we provided a model of how people can most effectively pursue their objectives when they are limited by a scarcity of means—there were no other people in the picture. Alexei, the student, and Angela, the farmer, cared about their own free time and their own grades or consumption. But when there are other people in the picture, people generally do not care only about what happens to themselves, but also what happens to others. When this is the case we say that the individual has *social preferences*. Altruism is an example of a social preference. Spite and envy are also social preferences.

Altruistic preferences as indifference curves

We can use the same feasible sets and indifference curves that you learned in Unit 3 to study how people interact when social preferences are part of people's motivations.

Choosing a point on your highest feasible indifference curve does not mean that you are self-interested, because *the indifference curve may represent altruistic preferences*.

To see this, imagine the following situation. Anil was given some tickets for the national lottery, and one of them won a prize of 10,000 rupees. He can, of course, keep all the money for himself, but he can also share some of it with his neighbour Bala. Figure 4.5 represents the situation graphically. The horizontal axis represents the amount of money Anil keeps for himself in thousands of rupees, and the vertical one the amount that he gives to Bala. Each point (x, y) represents a combination

of amounts of money for Anil (x) and Bala (y) in thousands of rupees. The shaded triangle depicts the feasible choices for Anil. At the corner (*10, 0*) on the horizontal axis, Anil keeps it all. At the other corner (*0, 10*) on the vertical axis, Anil gives it all to Bala. Anil's feasible set is the shaded area.

The boundary of the shaded area is the feasible frontier. If Anil is dividing up his prize money between himself and Bala, he chooses a point on that frontier (being inside the frontier would mean throwing away some of the money). The choice among points on the feasible frontier is called a *zero sum game* because, in choosing point *B* rather than point *A* in Figure 4.5, the sum of Anil's losses and Bala's gains is zero (for example, Anil has 3,000 fewer rupees at B than at A, and Bala has 3,000 rupees at *B* and nothing at *A*).

Anil's choice will be determined by his preferences, which can be represented by indifference curves, just as if he were choosing between goods and leisure in Unit 3. Here the indifference curve represents the combinations of how much Bala gets, and how much Anil keeps for himself, that are all equally preferred by Anil. In Figure 4.5 you can look at two cases: in the first Anil has self-interested preferences, in which case his indifference curves are straight vertical lines; in the second he is somewhat altruistic—he cares about Bala—and his indifference curves are downward-sloping.

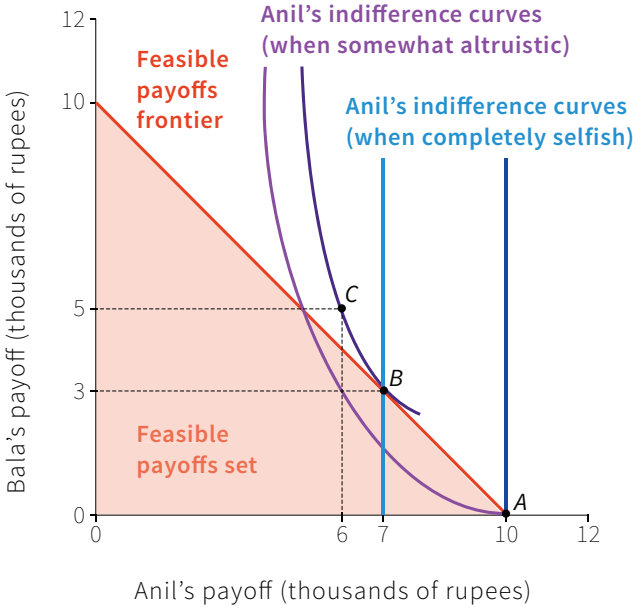


Figure 4.5 *How Anil chooses to distribute his lottery winnings depends on whether he is selfish or altruistic.*

If Anil is self-interested the best option, given his feasible set, is A, where he keeps all the money. If he *derives utility* from Bala's consumption, he has downward-sloping indifference curves and may prefer an outcome where Bala gets some of the money. With the specific indifference curves shown in Figure 4.5, the best feasible option for Anil is point B (*7, 3*) where Anil keeps 7,000 rupees and gives 3,000 to Bala. Anil prefers to give 3,000 rupees to Bala, even at a cost of 3,000 rupees to him. This is an

example of altruism: Anil is willing to bear a cost to benefit somebody else. If you are familiar with calculus, this Leibniz will show you how to find the best feasible option given Anil's altruistic utility function.

DISCUSS 4.3: ALTRUISM AND SELFLESSNESS

1. What would Anil's indifference curves look like if he cared just as much about Bala's consumption as his own?
2. What would they look like if he derived utility only from the total of his and Bala's consumption?
3. What would they look like if he derived utility only from Bala's consumption?
4. For each of these cases, provide an explanation of Anil's preferences.

4.5 ALTRUISTIC PREFERENCES IN THE PRISONERS' DILEMMA

Remember the prisoners' dilemma game in section 4.3 that showed how Anil and Bala were going to get rid of pests? It led to an unfortunate outcome partly because of the first problem we identified: Anil and Bala did not place any value on the payoffs of the other person, and so did not internalise the costs that their actions inflicted on the other.

We can now find out how altruistic preferences affect this.

In social interactions of this kind, there is a conflict between the unsatisfactory outcome that arises when both parties behave as self-interest dictates, and the outcome in which both are better off. The choice of pest control regime using the insecticide implied a *free ride* on the other farmer's contribution to ensuring clean water.

If Anil cares about Bala's wellbeing as well as his own, the outcome can be different. In Figure 4.6 the two axes now represent Anil and Bala's payoffs. Just as with the example of the lottery, the four points represent feasible outcomes. However, there are just four possible outcomes, rather than the set of feasible points in Figure 4.5. We have shortened the names of the strategies for convenience: *Terminator* is *T*, *IPC* is *I*. Notice that movements upward and to the right from (*T, T*) to (*I, I*) are win-win: both get higher payoffs. On the other hand, moving up, and to the left, or down, and

to the right—from (*I*, *T*) to (*T*, *I*) or the reverse—are win-lose changes. Win-lose means that Bala gets a higher payoff at the expense of Anil, or Anil benefits at the expense of Bala.

As in the case of dividing lottery winnings, we can look at two cases: if Anil does not care about Bala's wellbeing, his indifference curves are vertical lines; and if he does care, he has downward sloping indifference curves. Use the slideline to see what will happen in each case.

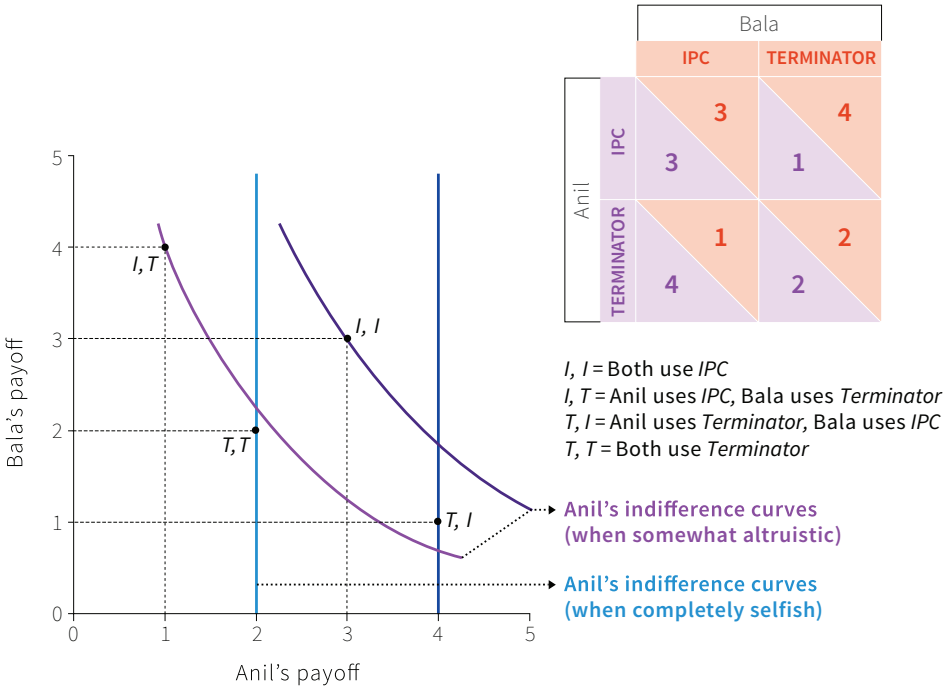


Figure 4.6 *Anil's decision to use IPC (I) or Terminator (T) as his crop management strategy depends on whether he is completely selfish or somewhat altruistic.*

If Anil does not care about Bala, he prefers (*T*, *I*) to (*I*, *I*), and he also prefers (*T*, *T*) to (*I*, *T*). So in that case Anil also prefers T to I whatever Bala chooses. T is unambiguously the best choice for Anil if he is completely selfish.

Things are different when Anil cares about Bala's wellbeing. In this case Anil has downward-sloping indifference curves, as shown in the figure. In that case, Anil prefers (*I*, *I*) to (*T*, *I*) and (*I*, *T*) to (*T*, *T*), so he prefers *I* to *T* whatever Bala chooses, which now makes using IPC unambiguously the better choice for Anil. If Bala feels the same way then the two would both choose *IPC*, resulting in the outcome that they both most prefer.

The main lesson is that *if people care about one another, social dilemmas are easier to resolve*. This helps us understand the historical examples in which people mutually cooperate for irrigation or enforce the Montreal Protocol to protect the ozone layer, rather than free riding on the cooperation of others.

DISCUSS 4.4: AMORAL SELF-INTEREST

Imagine a society in which everyone was entirely self-interested (cared only about his or her own wealth) and amoral (followed no ethical rules that would interfere with gaining that wealth). How would that society be different from the society you live in. Consider the following:

- Families
- Workplaces
- Neighbourhoods
- Traffic
- Political activity (would people vote?)

4.6 THE RULES OF THE GAME MATTER: PUBLIC GOODS AND PEER PUNISHMENT

Now let's look at the second reason for an unfortunate outcome in the prisoners' dilemma game: There was no way that either Anil or Bala (or anyone else) could make whoever used the insecticide pay for the harm that it caused.

The problems of Anil and Bala are hypothetical, but they capture the real dilemmas of free riding that face many people around the world. For example, as in Spain, many farmers in southeast Asia rely on a shared irrigation facility to produce their crops. The system requires constant maintenance and new investment. Each farmer faces the decision of how much to contribute to these activities. These activities benefit the entire community and, if the farmer does not volunteer to contribute, others may do the work anyway.

Imagine there are four farmers who are deciding whether to contribute to the maintenance of an irrigation project.

For each farmer, the cost of contributing to the project is $10. But when one farmer contributes, all four of them will benefit from an increase in their crop yields made possible by irrigation, so they will each gain $8. Contributing to the irrigation project is called a *public good*: when one individual bears a cost to provide the good, everyone receives a benefit.

Now, consider the decision facing Kim, one of the four farmers. Figure 4.7 shows how her total earnings depend on her decision, but also on the number of other farmers who decide to contribute to the irrigation project.

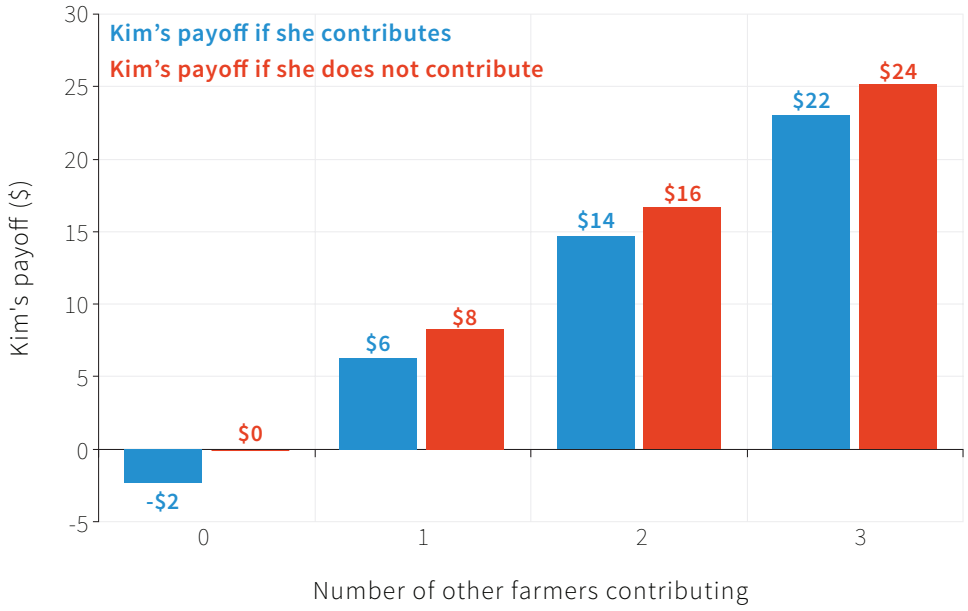


Figure 4.7 *Kim's payoffs in the public good game.*

For example, if two of the others contribute, Kim will receive a benefit of $8 from each of their contributions. So if she makes no contribution herself, her total payoff, shown in red, is $16. If she decides to contribute, she will receive an additional benefit of $8 (and so will the other three farmers). But she will incur a cost of $10, so her total payoff is $14, as in Figure 4.7, and as calculated in Figure 4.8.

Benefit from the contribution of others	16
Plus benefit from her own contribution	+8
Minus cost of her contribution	–10
Total	**$14**

Figure 4.8 *When two others contribute, Kim's payoff is lower if she contributes too.*

Figures 4.7 and 4.8 illustrate the social dilemma: whatever the other farmers decide to do, *Kim makes more money if she doesn't contribute than if she does*. She can free ride on the contributions of the others.

This public goods game is a prisoners' dilemma in which there are more than two players: if the farmers care only about their own monetary payoff, none would contribute and their payoffs would all be zero. On the other hand, if all contributed, each would get $22. Everyone benefits if everyone cooperates, but each would do better by free riding on the others irrespective of what others do. Therefore *Free ride* is a dominant strategy.

In the public goods game an uncontaminated water supply was a public good for the two farmers, Anil and Bala; if one chose not to use the pesticide, both were rewarded with clean water.

If large numbers of people are involved in a public goods game though, it is less likely that altruism will be sufficient to sustain a mutually beneficial outcome, such as neither using *Terminator* in the pest problem.

Repeated games

But, around the world, real farmers and fishing people have faced public goods situations in many cases with great success. The evidence gathered by Elinor Ostrom, a political scientist, and other researchers on common irrigation projects in India, Nepal, and other countries, shows that the degree of cooperation varies. In some communities a history of trust encourages cooperation. In others, cooperation does not happen. In south India, for example, villages with extreme inequalities of land and caste status had more conflicts over water usage. Less unequal villages maintained irrigation systems better: it was easier to sustain cooperation.

GREAT ECONOMISTS

ELINOR OSTROM

The choice of Elinor Ostrom (1933-2012), a political scientist, as a co-recipient of the 2009 Nobel Memorial Prize in Economics surprised most economists. For example, Steven Levitt, a professor at the University of Chicago, admitted he knew nothing about her work, and had "no recollection of ever seeing or hearing her name mentioned by an economist."

Some defended it. Vernon Smith, an experimental economist who had previously been awarded the Nobel prize, congratulated the Nobel committee for recognising her originality, "scientific common sense" and willingness to listen "carefully to data."

Ostrom's entire academic career was focused on a concept that plays a central role in economics but is seldom examined in much detail: *property*. Ronald Coase had established the importance of clearly delineated property rights when one person's actions affected the welfare of others. But Coase's main concern was the boundary between the individual and the state in regulating such actions. Ostrom explored the middle ground where communities, rather than individuals or formal governments, held property rights.

The conventional wisdom at the time was that informal collective ownership of resources would lead to a "tragedy of the commons". Thanks to Elinor Ostrom this is no longer a consensus view.

First, she made a distinction between resources held as common property and those subject to open access:

- Common property involves a well-defined community of users who are able in practice, if not under the law, to prevent outsiders from exploiting the resource: inshore fisheries, grazing lands, or forest areas are examples.
- Open access resources can be exploited without restrictions, other than those imposed by states: ocean fisheries or the atmosphere as a carbon sink, for instance.

Ostrom was not alone in stressing this distinction, but she drew on a unique combination of case studies, statistical methods, game theoretic models with unorthodox ingredients, and laboratory experiments to try to understand how tragedies of the commons could be averted.

Ostrom discovered great diversity in how common property is managed. Some communities were able to devise rules and draw on social norms to enforce sustainable resource use, while others failed to do so. She spent much of her career trying to identify what determined success, and use theory to understand why.

Many economists believed that the diversity of outcomes could be understood using the theory of repeated games, which predicts that, even when all individuals care only for themselves, if interactions are repeated with sufficiently high likelihood, and individuals are patient enough, then cooperative outcomes can be sustained in equilibrium.

But this was not a satisfying explanation for Ostrom, partly because the same result predicted that any outcome, including rapid depletion, could also be an equilibrium.

More importantly, Ostrom knew that sustainable use was enforced by actions that clearly deviated from the hypothesis of material self-interest. In particular, individuals would willingly bear considerable costs to punish violators of rules or norms. Paul Romer, an economist, said she recognised the need to "expand models of human preferences to include a *contingent taste for punishing others*."

Ostrom developed simple game theoretic models in which individuals have unorthodox preferences, caring directly about trust and reciprocity. And she looked for the ways in which people faced with a social dilemma avoided tragedy by changing the rules so that the strategic nature of the interaction was transformed.

She worked with economists to run a pioneering series of experiments, confirming the widespread use of costly punishment in response to excessive resource extraction, and also demonstrated the power of communication and the critical role of informal agreements in supporting cooperation. Thomas Hobbes, a 17th century philosopher, had asserted that agreements had to be enforced by governments. He stated "Covenants [agreements] without the sword are but words." Ostrom disagreed. As she wrote in the title of an influential article, covenants—even without a sword—make self-governance possible.

Social preferences partly explain why these communities avoid Garrett Hardin's *Tragedy of the commons*. More importantly, the communities found ways to change the rules of the game so that they were no longer imprisoned in the simple one-shot game with a mutually disadvantageous dominant strategy equilibrium (the game that resulted in Anil and Bala's use of *Terminator*). In many cases the one-shot prisoners' dilemma is not an adequate description of the problems that these communities face.

This is an important feature of social interactions: *life is not a one-shot game*.

Free riding on the contributions of others today may have unpleasant consequences for the free rider tomorrow or years from now. In game theory, when the same interaction takes place again and again, we call this a *repeated game*.

The interaction between Anil and Bala in our model was a one-shot game. But as owners of neighbouring fields, Anil and Bala are more realistically portrayed as engaged in a repeated game.

Imagine how differently things would work out if we represent their interaction as a repeated game. Suppose that Bala has adopted *IPC*; what is Anil's best response? He would reason like this:

Anil If I play *IPC* then maybe Bala will continue to do so, but if I use *Terminator*—which would raise my profits this season—Bala would use *Terminator* next year. So unless I am extremely impatient for income now, I'd better stick with *IPC*.

Bala could reason in exactly the same way. The result might be that they would then continue playing *IPC* for ever.

The fact that social interactions like this will continue in the future also can sustain high levels of cooperation in a public goods game, as long as people have opportunities to target free riders once it becomes clear who is contributing less than the norm.

To see how this works, here is an experiment about contributions in a public goods game. In this experiment people have opportunities to punish free riders.

Figure 4.9a shows the results of laboratory experiments that mimic the costs and benefits from contribution to a public good in the real world. The experiments were conducted in cities around the world. In each experiment participants play 10 rounds of a public goods game, similar to the one involving Kim and the other farmers that we just described. In each round, the people in the experiment (here we call them *subjects*) are given $20. They are randomly sorted into small groups, typically of four people, who don't know each other. They are asked to decide on a contribution from their $20 to a common pool of money. The pool is a public good: for every dollar contributed, each person in the group, including the contributor, receives $0.40.

Imagine that you are playing the game, and you expect the other three members of your group each to contribute $10. Then if you don't contribute you will get $32 (three returns of $4 from their contributions, plus the initial $20 that you keep). The others have paid $10, so they only get $32 - $10 = $22 each. On the other hand, if you also contribute $10, then everyone, including you, will get $22 + $4 = $26. Unfortunately for the group, you do better by not contributing—that is, because the reward for free riding ($32) is greater than for contributing ($26). And, unfortunately for you, the same applies to each of the other members.

After each round, the participants can see the total amount contributed, but not the amount that each of the others has contributed. In Figure 4.9a, each line represents the evolution over time of average contributions in a different location around the world. Just as in the prisoners' dilemma, people are definitely not solely self-interested.

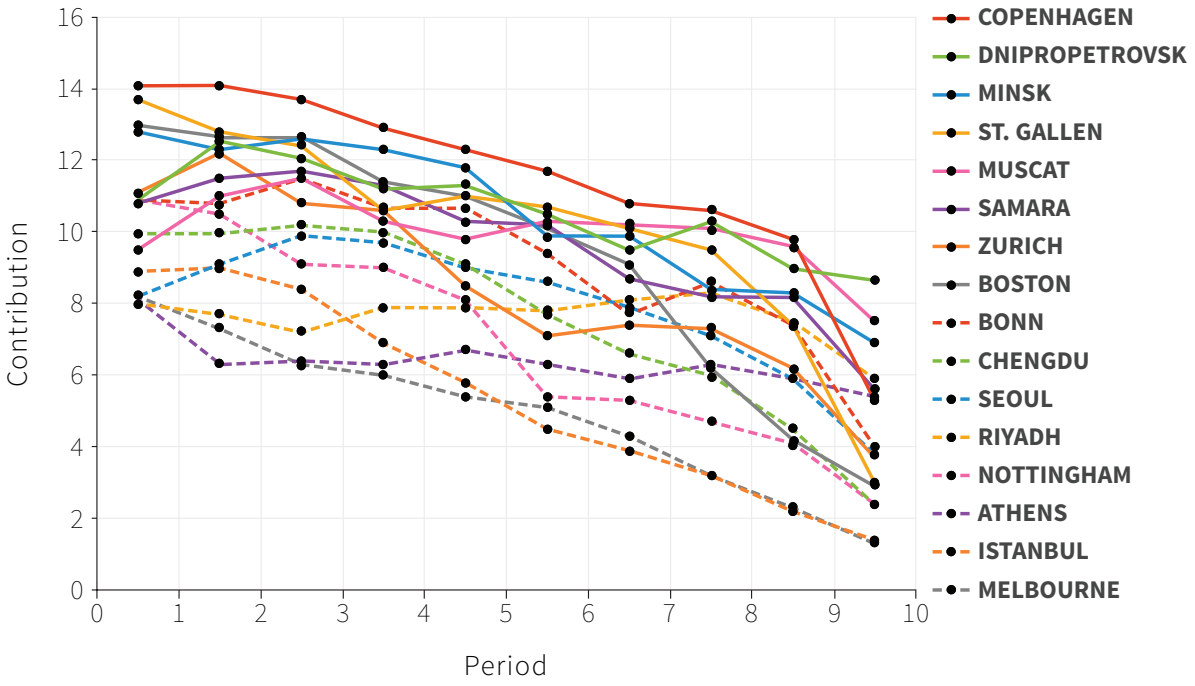


Figure 4.9a *Worldwide public good experiments: Contributions over 10 periods.*

Source: *Figure 3 from Herrmann, Benedikt, Christian Thoni, and Simon Gachter. 2008. 'Antisocial Punishment Across Societies.' Science 319 (5868): 1362–67.*

As you can see from Figure 4.9a, players in Chengdu contributed $10 in the first round, just as we described above. In every population where the game is played, contributions to the public good are high in the first period, although much more so in some cities (Copenhagen) than in others (Melbourne). This is remarkable: if you care only about your own payoff, contributing *nothing at all* is the dominant strategy. The high initial contributions could have occurred because the participants in the experiment valued their contribution to the payoffs that others received (they were altruistic). But the difficulty (or, as Hardin would have described it, the tragedy) is obvious: everywhere, the contributions to the public good decreased over time.

Nevertheless, the results also show that despite a large variation across societies, most of them still have high contribution levels at the end of the experiment.

The most plausible explanation of the pattern is *not* altruism. It is likely that contributors decrease their level of cooperation if they observe that others are contributing less than expected, and therefore free riding on them. It seems as if those contributing more than the average would like to punish the low contributors for their unfairness, or for violating a *social norm* of contributing. The

SOCIAL NORM

- An understanding that is common to most members of a society about what people should do in a given situation when their actions affect others.

last thing they want to do is to increase the payoffs of free riders by contributing more to the public good. The only way to punish free riders in this experiment is to stop contributing. This is the tragedy of the commons.

Many people are happy to contribute as long as others reciprocate. A disappointed expectation of reciprocity is the most convincing reason that contributions fall so regularly in later rounds of this game.

To test this, the experimenters took the public good game experiment shown in Figure 4.9a and introduced a punishment option, the results of which are shown in Figure 4.9b. For the majority of subjects, including those in China, South Korea, northern Europe and the English-speaking countries, contributions increased when they had the opportunity to punish free riders.

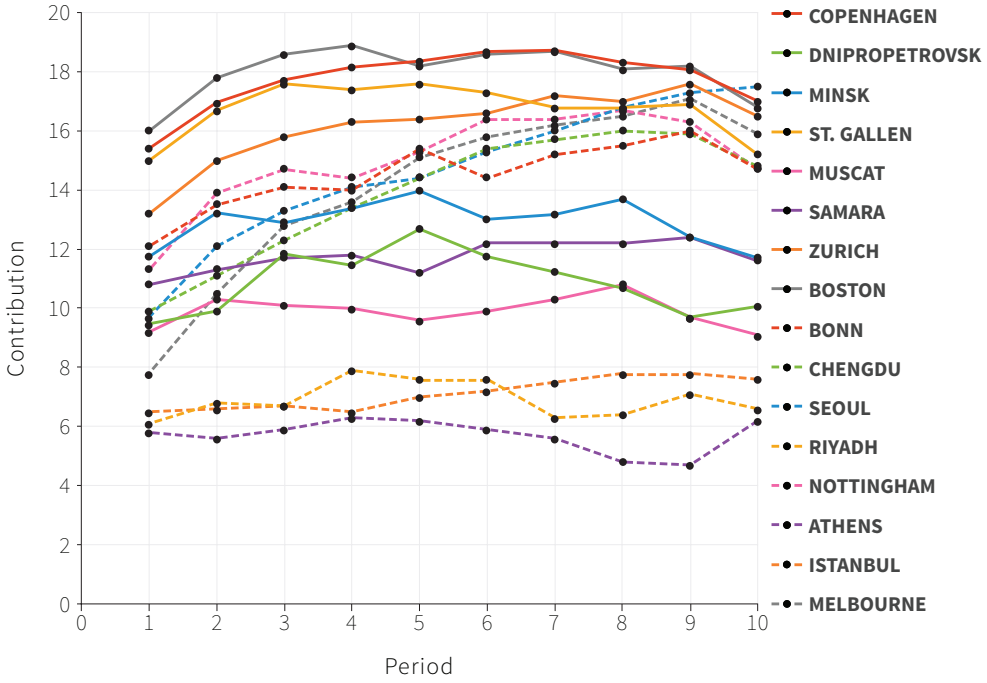


Figure 4.9b *Worldwide public good experiments: With opportunities for peer punishment.*

Source: *Figure 3 from Herrmann, Benedikt, Christian Thoni, and Simon Gachter. 2008. 'Antisocial Punishment Across Societies.' Science 319 (5868): 1362–67.*

People who consider that others have been unfair, or have violated a social norm, may retaliate—even if the cost to themselves is high. Their punishment of others is a form of altruism, because it costs them something to help deter free riding behaviour that is detrimental to the wellbeing of most members of the group.

This experiment illustrates the way that, even in large groups of people, a combination of repeated interactions and social preferences can support high levels of contribution to the public good. This can occur even when a one-shot public goods game, or a game with no opportunities for punishing free riders, would lead to a different result.

The public goods game, like the prisoners dilemma, is a situation in which there is something to gain for everyone by engaging with others in a common project such as pest control, maintaining an irrigation system or controlling carbon emissions. But there is also something to lose, when others free ride.

Cooperation

Cooperation means participating in a common project in such a way that mutual benefits occur. Cooperation need not be based on an agreement. To see this, recall the games we have studied:

- *The invisible hand*: Anil and Bala acted entirely independently, but the division of labour that results from their pursuit of their own interests also results in mutual gains. Neither could do better by adopting another strategy. Their engagement in the village market facilitates this kind of cooperation without agreements.
- *The prisoners' dilemma*: If their pest control interaction is repeated, they could refrain from using *Terminator* simply by individually working out the future losses they would suffer as a result of abandoning IPC.
- *The public goods game*: With punishment of free riders, the payers made no agreements about how to play, but their willingness to punish others sustained high levels of cooperation in many countries.

But, as we will see, cooperation sometimes breaks down because of conflicts of interest over how the mutual gains to cooperation will be shared.

DISCUSS 4.5: ARE LAB EXPERIMENTS ALWAYS VALID?

In 2007 Steven Levitt and John List published a paper called "What Do Laboratory Experiments Measuring Social Preferences Reveal about the Real World?". (Click on the link to download it. You will also find the reference listed in the Read more section at the end of this unit.)

1. Why, and how, might people's behaviour in real life vary from what has been observed in laboratory experiments?
2. Using the example of the public goods experiment in this section, explain why you might observe systematic differences between the observations recorded in Figures 4.9a and 4.9b, and what might happen in real life.

BEHAVIOURAL EXPERIMENTS IN THE LAB AND IN THE FIELD

To understand economic behaviour we need to know about people's preferences. In the previous unit, for example, students and farmers valued free time. How much they valued it was part of the information we needed to predict how much time they spend studying and farming.

In the, past economists have learned about our preferences from:

- *Survey questions*: To determine political preferences, brand loyalty, degree of trust of others, or religious orientation.
- *Statistical studies of economic behaviour*: For example purchases of one or more goods when the relative price varies—to determine preferences for the goods in question. One strategy is to reverse engineer what the preferences must have been, as revealed by purchases. This is called revealed preference.
- *Behavioural experiments*: This method has become important in the empirical study of preferences. Part of the motivation for experiments is that understanding peoples' altruism, reciprocity, inequality aversion as well as self-interest is essential to being able to predict how they will behave as employees, family members, custodians of the environment, and citizens.

Surveys have a problem: asking someone if they like ice cream will probably get an honest answer. But the answer to the question: "How altruistic are you?" may be a mixture of truth, self-advertising, and wishful thinking. This is why economists sometimes use experiments to discover our preferences. Statistical studies cannot control the situation in which the preferences were revealed, so they make it difficult to compare different groups.

Experiments do not measure what people say, but what they do. Experiments are designed to be as realistic as practical, while controlling the situation:

- *Decisions have consequences*: The decisions in the experiment may decide how much money the subjects earn by taking part. Sometimes the stakes can be as high as a month's income.
- *Instructions, incentives and rules are common to all subjects*: There is also a common *treatment*. This means that, if we want to compare two groups, the only difference between the control and treatment groups is the treatment itself, so that its effects can be identified.
- *Experiments can be replicated*: They are designed to be implemented with other groups of participants.
- *Experimenters attempt to control for the variables of interest*: Other variables, as far as possible, are kept constant, because they may affect the behaviour we want to measure.

This means that when people behave differently in the experiment it provides evidence about differences in their preferences, not in the situation that each person faces.

Economists have studied public goods extensively using laboratory experiments in which the subjects are asked to make decisions about how much to contribute to a public good. In some cases economists have designed experiments that closely mimic real-world social dilemmas. The work of Juan Camilo Cárdenas, an economist at the Universidad de los Andes in Bogotá, Colombia is an example. He performs experiments about social dilemmas with people who are facing similar problems in their real life, such as overexploitation of a forest or of a fish stock. In this video he describes his use of experimental economics in real-life situations, and how it helps us understand why people cooperate—even when there are apparent incentives not to do so.

Economists have discovered that the way people behave in experiments can be used to predict how they react in other situations. For example, fishermen in Brazil who acted more cooperatively in an experimental game also practiced fishing in a more sustainable manner than the fishermen who were less cooperative in the experiment.

For a summary of the kinds of experiments that have been run, the main results, and whether behaviour in the experimental lab predicts behaviour in other arenas, read the research done by some of the economists who specialise in experimental economics: for example, Colin Camerer and Ernst Fehr, Armin Falk and James Heckman, or the experiments done by Joseph Heinrich and a large team of collaborators around the world.

Stephen Levitt and John List, however, raise concerns about *external validity*: do people behave the same way in the street as they do in the lab?

4.7 CONFLICTS OF INTEREST AND SOCIAL NORMS

There is a third aspect of the interaction that created an unfortunate outcome: Anil and Bala were not able to make an agreement about what each would do. Had they been able to, they could have simply agreed to use *IPC* or legislated a ban on the use of *Terminator*.

People commonly resort to negotiation to solve their economic and social problems, but they do not always succeed. Consider, for example, a professor who might be willing to hire a student as a research assistant for the summer. In principle, both have something to gain from the relationship, because this might also be a good opportunity for the student to earn some money and learn. In spite of the potential for mutual benefit, there is also some room for conflict. The professor may want to pay less and have more of his research grant left over to buy a new computer, or he may need the work to be done quickly, meaning the student can't take time off. After negotiating, they may reach a compromise and agree that the student can earn a small salary while working from the beach. Or, perhaps, the negotiation will fail.

There are many situations like this in economics. A negotiation (sometimes called *bargaining*) is also an integral part of politics, foreign affairs, law, social life and even family dynamics. A parent may give a child a smartphone to play with in exchange for a quiet evening; a country might consider giving up land in exchange for peace; a government might be willing to negotiate a deal with student protesters to avoid political instability. As with the student and the professor, each of these bargains might not be struck: maybe they wouldn't be willing to do these things.

When do negotiations succeed?

To help think about what makes a deal work, consider the following situation. You and a friend are walking down the street and you see a $100 note on the ground. How would you decide how to split your lucky find? If you split the amount equally, this could be described as reflecting a *social norm* in your community that says that something you get by luck should be split 50-50.

Dividing something of value in equal shares (the 50-50 rule) is a social norm in many communities, as is giving gifts on birthdays to close family members and friends. Social norms are common to an entire group of people (almost all follow them) and tell a person what they should do in the eyes of most people in the community.

Preferences include norms, but they also include many other "pro" and "con" attitudes that are reflected in behaviour:

- *Preferences need not be about what one should do*: You can like ice cream without thinking that everyone (or even you) should enjoy eating it.
- *Preferences typically differ from person to person*: Recall that norms are ideas about social behaviour that are common to an entire group. People in the same group, though, can have different preferences (you may like ice cream, but maybe your friend hates it).

We would expect that, even if there were a 50-50 norm in a community, some individuals might not respect the norm exactly. Perhaps some people act more selfishly than the norm requires and some more generously. What happens next will

depend both on the social norm (a fact about the world, and which reflects attitudes to fairness that have evolved over long periods), but also on the preferences of the individuals concerned.

Suppose the person who saw the money first has picked it up. There are at least three reasons why that person might give some of it to a friend:

- *Altruism*: We have already considered the first, in the case of Anil and Bala. This person might be altruistic and care about the other being happy, or about some other aspect of the other's wellbeing.
- *Fairness*: Or, the person holding the money might think that 50-50 is fair. In this case, the person is motivated by fairness, or what economists term inequality aversion.
- *Reciprocity*: The friend may have been kind to the lucky money-finder in the past, or kind to others, and deserves to be treated generously because of this. In this case we say that our money-finder has reciprocal preferences.

These social preferences all influence our behaviour, sometimes working in opposite directions. This would be the case when the money-finder has strong fairness preferences, but knows that the friend is entirely selfish. The fairness preferences tempt the finder to share; the reciprocity preferences push the finder to keep the money.

4.8 DIVIDING A PIE (OR LEAVING IT ON THE TABLE)

One of the most common tools to study social preferences is the two-person one-shot game known as the *ultimatum game*. It has been used all around the world with experimental subjects that have included students, farmers, warehouse workers, and hunter-gatherers. Experiments using this game allow us to investigate how economic outcomes, in this case how something of value will be divided, depend on individual preferences such as pure self-interest, altruism, inequality aversion, and reciprocity.

In the experiment, a group of people (the subjects of the experiment) are invited to play a game in which they will win some money. How much they win will depend on how they and the others in the game play. Real money is at stake in experimental games like these because, unless real money were on the table, we could not be sure the subject's answers to a hypothetical question would reflect their actions in real life.

The rules of the game are explained to the players. There are two roles in the game, a *Proposer* and a *Responder*, assigned at random. The subjects do not know each other, but they know the other player was recruited to the experiment in the same way. Subjects remain anonymous.

The Proposer is provisionally given an amount of money, say $100, by the experimenter, who instructs the Proposer to offer the Responder part of it. Any split is permitted, including keeping it all, or giving it all away. We will call this amount the "pie" because the point of the experiment is how it will be divided up.

The split takes the form: "*x* for me, *y* for you" where $x + y = \$100$. The Responder knows that the Proposer has $100 to split. After observing the offer, the Responder accepts or rejects it. If the offer is rejected, both individuals get nothing. Otherwise, if the offer is accepted, the split is implemented and the Proposer gets x and the Responder y. For example, if the Proposer offers $35 and the Responder accepts the offer, the Proposer gets $65 and the Responder gets $35. If the Responder rejects the offer, they both get nothing.

This is called a *take-it-or-leave-it offer*. It is the ultimatum in the game's name. The Responder is faced with a choice: accept $35, or get nothing.

This is a game about dividing up the *economic rents* that arise in an interaction. The slice of the pie that each of the two players receives is a rent, because it is what they get above their next best alternative (which, in this case, is to get nothing). In Unit 2 we saw that entrepreneurs who were the first to introduce a new technology get an innovation rent, that is, profits greater than would have been possible without the new technology. In the experiment the rent arises because the experimenter provisionally gives the Proposer the pie to divide. In the ultimatum game example above, if the Responder accepts the Proposer's offer, then the Proposer gets a rent of $65, and the Responder gets $35.

If the Responder rejects the offer, however, they both get no rent at all (essentially they throw away the pie). For the Responder there is a cost to saying no. He loses the rent that he would have received. The Proposer's offer of $35 is therefore the opportunity cost of rejecting the offer.

In Figure 4.10 the ultimatum game is shown in a diagram called a game tree. Figure 4.10 shows a simplified setup in which the proposer's choices are either the "fair offer" of an equal split, (5, 5) in the figure, or the "unfair offer" of 2 (keeping 8 for herself). Then the respondent has the choice to accept or reject. The payoffs are shown in the last row.

The game tree is a useful way to represent social interactions because it clarifies who does what, when, and what are the results. The game tree for the ultimatum game makes it clear that this game differs from the previous games because here one player

(the Proposer) chooses her strategy first, followed by the Responder. Previously we assumed that players chose strategies simultaneously. This is called a *sequential game* (not surprisingly, the earlier games are called *simultaneous games*).

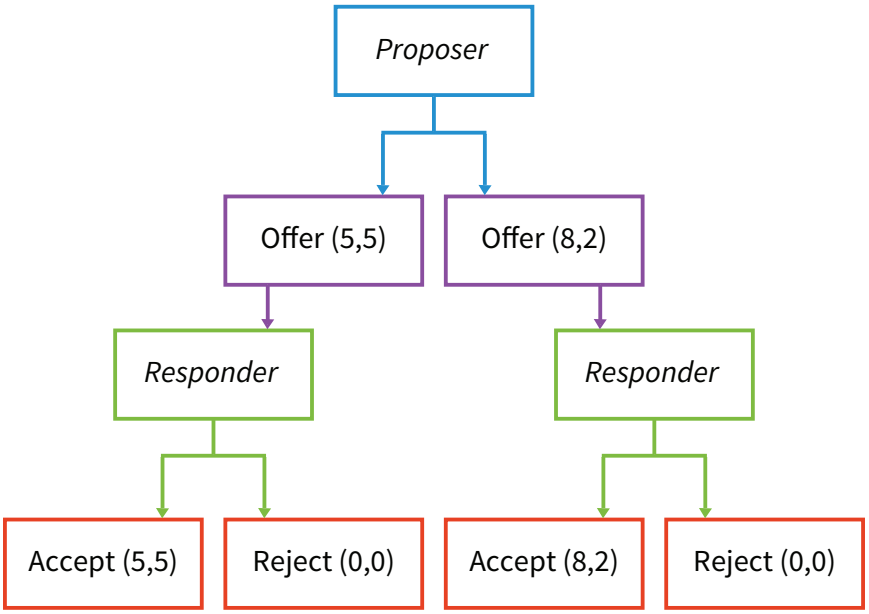


Figure 4.10 *Game tree for the ultimatum game.*

What the Proposer will get depends on what the Responder does, so the Proposer has to think about the likely response of the other player. That is why this is called a strategic interaction. If you're the Proposer you can't try out a low offer to see what happens: you have only one chance at making the offer.

Put yourself in the place of the Responder in this game. What is the minimum offer you are willing to accept? Now switch roles. Suppose that you are the Proposer. What split would you offer to the Responder? Would your answer depend on whether the other person was a friend, a stranger, a person in need, or a competitor?

A Responder who thinks that the Proposer's offer has violated a social norm of 50-50, or that for some other reason that the offer is insultingly low, might be willing to sacrifice the payoff to punish the Proposer.

If you work through this unit's Einstein, and Discuss 4.11 that follows it, you will see how to work out the *minimum acceptable offer*, taking account of the social norm and of the individual's own attitude to reciprocity. The minimal acceptable offer is the offer at which the pleasure of getting the money is equal to the satisfaction the person would get from refusing the offer and getting no money, but punishing the Proposer for violating the social norm of 50-50. If you are the Responder and your minimum acceptable offer is $35 (of the total pie of $100) then, if the Proposer offered you $36, you might not like the Proposer much; but this violation of the 50-50 norm

would not motivate you to punish the Proposer by rejecting the offer. If you did that, you would go home with satisfaction worth $35 and no money, when you could have had $36 in cash.

4.9 FAIR FARMERS, SELF-INTERESTED STUDENTS?

If you are a Responder who cares only about your own payoffs, you should accept any positive offer because something, no matter how small, is always better than nothing. Therefore, in a world composed only of self-interested individuals, the Proposer would anticipate that the Responder would accept any offer and, for that reason, would then offer the minimum possible amount: one cent, knowing it would be accepted.

Does this prediction match the experimental data? No, it does not. Just as with the prisoners' dilemma, we don't see the outcome we would predict if people were entirely self-interested. One-cent offers get rejected.

To see how farmers in Kenya and students in the US played this game, look at Figure 4.11. The height of each bar indicates the fraction of Responders who were willing to accept the offer indicated on the horizontal axis. Offers of more than half of the pie were acceptable to all of the subjects in both countries, as you would expect.

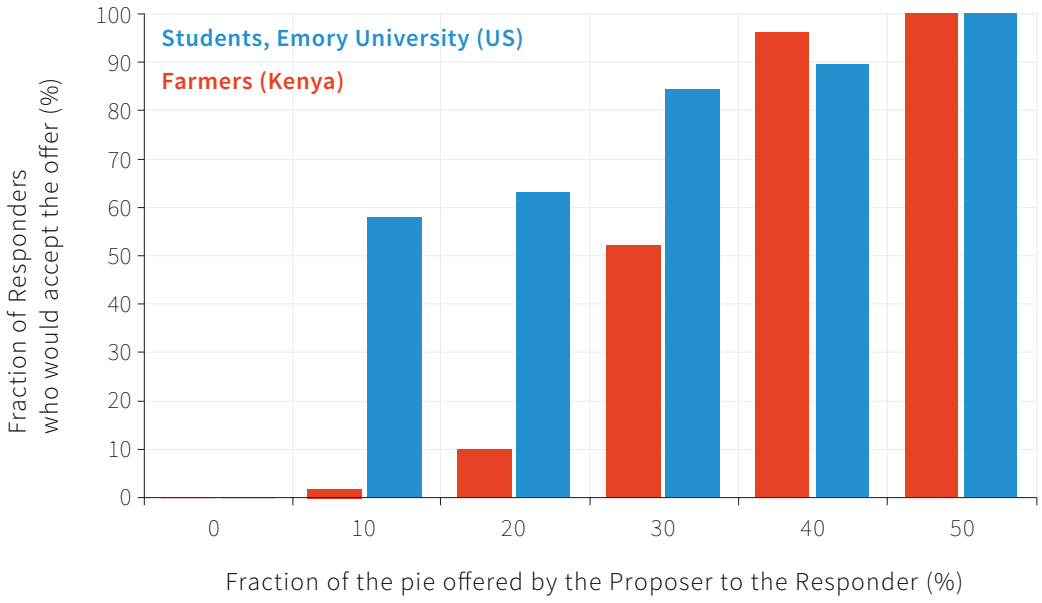


Figure 4.11 *Acceptable offers in the ultimatum game.*

Source: *Adapted from Adapted from Henrich, Joseph, Richard McElreath, Abigail Barr, Jean Ensminger, Clark Barrett, Alexander Bolyanatz, Juan Camilo Cardenas, et al. 2006. 'Costly Punishment Across Human Societies.' Science 312 (5781): 1767–70.*

Notice that the Kenyan farmers are very unwilling to accept low offers, presumably regarding them as unfair, while the US students are much more willing to accept low offers. For example, virtually all (90%) of the farmers would say no to an offer of one-fifth of the pie (the Proposer keeping 80%), while among the students, 63% would agree to such a low offer. More than half of the students would accept an offer of just 10% of the pie; almost none of the farmers would.

Although the results in Figure 4.11 indicate that attitudes differ towards what is fair, and how important fairness is, nobody in the Kenyan and US experiments was willing to accept an offer of zero, even though by rejecting it they would also receive zero.

DISCUSS 4.6: SOCIAL PREFERENCES

1. Which of the social preferences discussed above do you think motivated the subjects' willingness to reject low offers, even though by doing so they would receive nothing at all?
2. Why do you think that the Kenyan farmers were different from the US students?
3. Play the game described in this section using two separate sets of players, first your classmates and then your family and friends outside class. Is there any difference in the responses of these two groups? Explain.

The full height of each bar in Figure 4.12 indicates the percentage of the Kenyan and American Proposers who made the offer shown on the horizontal axis. For example, half of the farmers made proposals of 40%. Another 10% offered an even split. Among the students, shown in blue, only 11% made such generous offers.

But were the farmers really generous? To answer you have to think not only about how much they were offering, but also what they must have reasoned when considering whether the Respondent would accept the offer. If you look at Figure 4.12 and concentrate on the Kenyan farmers, you will see that very few proposed to keep the entire pie by offering zero (4% of them as shown in the far left-hand bar) and all of those offers would have been rejected (the entire bar is dark).

On the other hand, looking at the far right of the figure, we see that in the case of the Kenyan farmers, making an offer of half the pie ensured an acceptance rate of 100% (the entire bar is light). Those who offered 30% were about equally likely to see their offer rejected as accepted (the dark part of the bar is nearly as big as the light part).

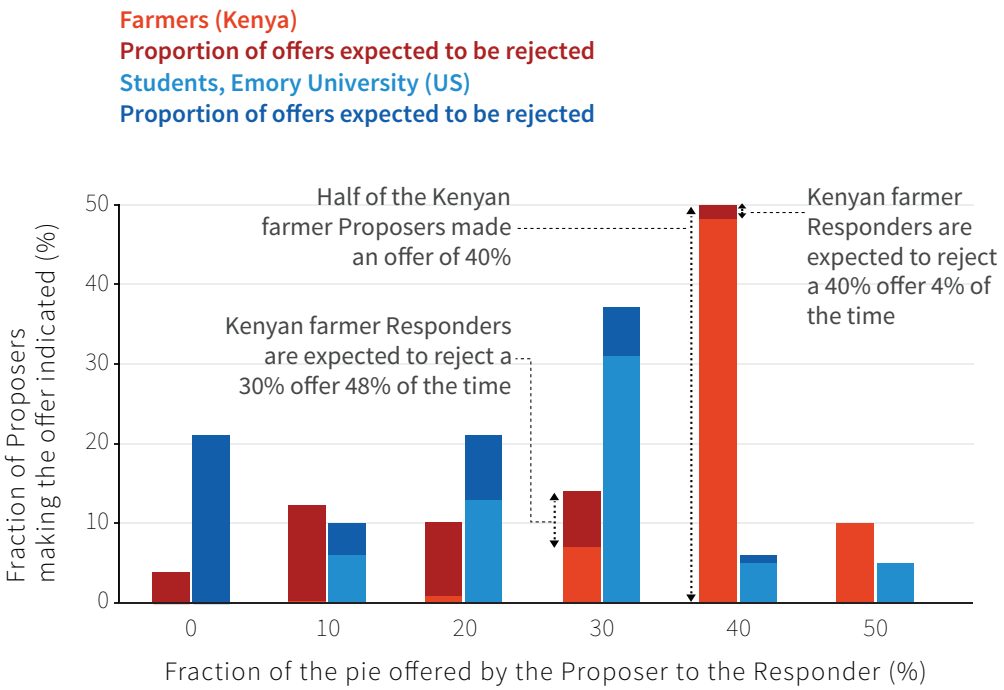


Figure 4.12 *Actual offers in the ultimatum game and expected rejections.*

Source: *Adapted from Henrich, Joseph, Richard McElreath, Abigail Barr, Jean Ensminger, Clark Barrett, Alexander Bolyanatz, Juan Camilo Cardenas, et al. 2006. 'Costly Punishment Across Human Societies.' Science 312 (5781): 1767–70.*

A Proposer who wanted to earn as much as possible would choose something between the extreme of trying to take it all or dividing it equally. The farmers who offered 40% were very likely to see their offer accepted and receive 60% of the pie. In the experiment, half of the farmers chose an offer of 40%. We would expect the offer to be rejected only 4% of the time, as can be seen from the dark-shaded part of the bar at the 40% offer in Figure 4.12.

Now suppose you are a Kenyan farmer and all you care about is your own payoff. Offering to give the Responder nothing is out of the question because that will ensure that you get nothing when they reject your offer. Offering half will get you half for sure—because the respondent will surely accept.

But you suspect that you can do better.

A proposer who cares only about his own payoffs will compare what is called the expected payoffs of the two offers: that is, the payoff that one may expect, given what the other person is likely to do (accept or reject) in case this offer is made. Your expected payoff is the payoff you get if the offer is accepted, multiplied by the probability that it will be accepted (remember that if the offer is rejected, the Proposer gets nothing). Here is how the Proposer would calculate the expected payoffs of offering 40% or 30%:

Expected payoff of offering 40%:
= 96% chance of keeping 60% of the pie
= 0.96 x 0.60
= 58%

Expected payoff of offering 30%:
= 52% chance of keeping 70% of the pie
= 0.52 x 0.70
= 36%

We cannot know if the farmers actually made this calculation, of course. But if they did they would have discovered that offering 40% maximised their expected payoff. This contrasts with the case of the acceptable offers in which considerations of inequity aversion, reciprocity or the desire to uphold a social norm were apparently at work. Unlike the Responders, many of the Proposers may have been trying to make as much money as possible in the experiment and had guessed correctly what the Responders would do.

Similar calculations indicate that, among the students, the expected payoff-maximising offer was 30%, and this was the most common offer among them. The students' lower offers could be because they correctly anticipated that lowball offers (even as low as 10%) would sometimes be accepted. They may have been trying to maximise their payoffs and hoped that they could get away with making low offers.

DISCUSS 4.7: OFFERS IN THE ULTIMATUM GAME

1. Why do you think that some of the farmers offered more than 40%? Why did some of the students offer more than 30%?
2. Why did some offer less?
3. Which of the social preferences that you have studied might have been involved?

How do the two populations differ? Many of both the farmers and the students offered an amount that would maximise their expected payoffs. The similarity ends there. The Kenyan farmers were more likely to reject low offers. Is this a difference between Kenyans and Americans, or between farmers and students? Or is it something unrelated to nationality and occupation entirely, but reflecting a local social norm? Experiments alone cannot answer these interesting questions; but before you jump to the conclusion that Kenyans are more averse to unfairness than

Americans, when the same experiment was run with rural Missourians in the US, they were even more likely to reject low offers than the Kenyan farmers. Almost every Missourian Proposer offered half of the pie.

DISCUSS 4.8: STRIKES

A strike over pay or working conditions may be considered an example of an ultimatum game.

1. Research a well-known strike and explain how it satisfies the definition of an ultimatum game.
2. Draw a game tree to represent this situation.
3. In this section, you have been presented with experimental data to test the predictions of the ultimatum game. How would you use the data from the strike you have researched to do the same?

4.10 THE RULES OF THE GAME MATTER: COMPETITION IN THE ULTIMATUM GAME

Social preferences provide a way of explaining why behaviour in ultimatum games departs from what purely self-interested individuals might do. But, as usual, things can be more complicated. For example, the professor looking for a research assistant could consider several applicants rather than just one. In this case, one would expect that negotiations would be affected by competition.

To think about the implications of increased competition, imagine a new ultimatum game in which a Proposer offers a two-way split of $100 to two respondents, instead of just one. In this version of the game, if either of the Responders accepts but not the other, that Responder and the Proposer get the split, and the other Responder gets nothing. If no one accepts, no one gets anything, including the Proposer. If both Responders accept, one is chosen at random to receive the split.

If you are one of the Responders, what is the minimum offer you would accept? Are your answers any different, compared to the original ultimatum game with a single Responder? Perhaps. If I knew that my fellow competitor is strongly driven by 50-

50 split norms, my answer would not be too different. But what if I suspect that my competitor wants the reward very much, or does not care too much about how fair the offer is?

And, suppose you are the Proposer now. What split would you offer?

Figure 4.13 shows laboratory evidence for the ultimatum game when there are two Responders playing multiple rounds. It's important to know that the participants are anonymous (why?).

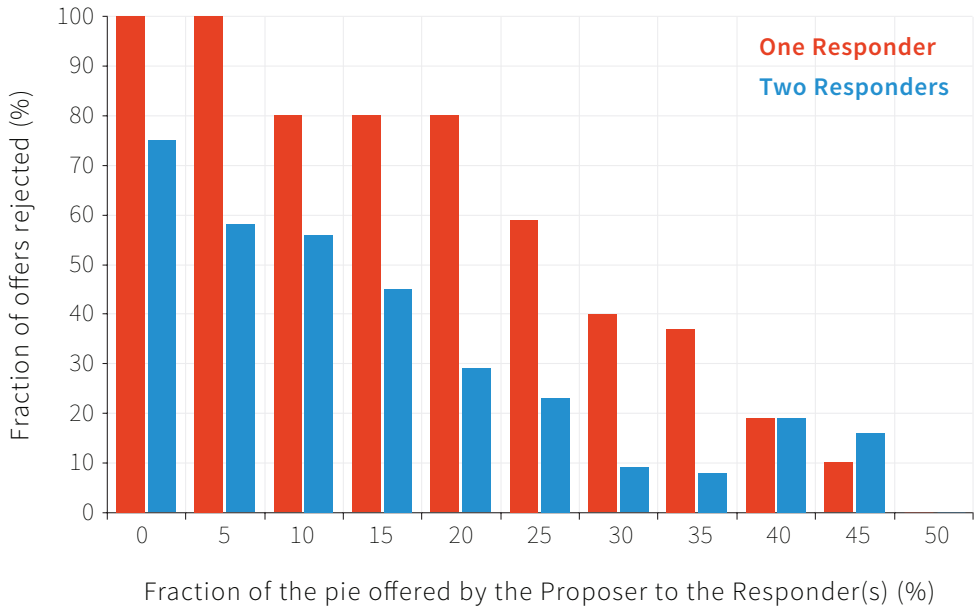


Figure 4.13. *Fraction of offers rejected by offer size in the ultimatum game with one and two Responders.*

Source: *Adapted from Figure 6 in Fischbacher, Urs, Christina M. Fong, and Ernst Fehr. 2009. 'Fairness, Errors and the Power of Competition.' Journal of Economic Behavior & Organization 72 (1): 527–45.*

The red bars show the fraction of offers that are rejected when there is a single Responder. The blue bars show the behaviour in experiments with two Responders. It is clear that competition among the Responders moves the observations closer to what we would see in a world populated by self-interested individuals who are concerned mostly about their own monetary payoffs.

To explain this phenomenon to yourself, think about what happens when a Responder rejects a low offer: this means getting a zero payoff. Unlike the situation in which there is a sole Responder, the Responder in a competitive situation *cannot be sure the Proposer will be punished*, because the other Responder may accept the low offer (not everyone has the same norms about proposals, or is in the same state of need).

Consequently, even fair-minded people will accept low offers to avoid having the worst of both worlds. Of course, the Proposers also know this, so they will make lower offers, which Responders still accept. Notice how a small change in the rules or the situation can have a big effect on the outcome. As in the public goods game where the addition of an option to punish free riders greatly increased the levels of contribution, changes in the rules of the game matter.

DISCUSS 4.9: A SEQUENTIAL PRISONERS' DILEMMA

Return to the prisoners' dilemma pest control game that Anil and Bala played, but now suppose that the game is played sequentially, like the ultimatum game. One player (chosen randomly) chooses a strategy first (the first mover), and then the second moves (the second mover).

1. Suppose you were the second mover and that the first mover had chosen *IPC*. What would you choose?
2. Suppose you are first mover and you know that the second mover has strong *reciprocal preferences*, meaning the second mover will act kindly towards someone who upholds social norms. What would you do?

4.11 SOCIAL INTERACTIONS: CONFLICTS IN THE CHOICE AMONG NASH EQUILIBRIA

In the invisible hand game, the prisoners' dilemma, and the public goods game, the action that gave a player his highest payoffs did not depend on what the other player did: there was a dominant strategy for each player, and hence a single dominant strategy equilibrium.

But this is often not the case.

We mentioned in passing a situation in which it is definitely untrue: driving on the right or driving on the left. Which side you drive on will depend on which country you are in: drive on the right in the US because you expect other drivers to drive on the right there, but drive on the left in Japan because you expect other drivers to do that too.

Both driving on the right in the US, and driving on the left in Japan, are termed *Nash equilibria*.

There are two Nash equilibria in the driving game because the best response to everyone else driving on the left is to drive on the left and similarly, when everyone is driving towards you on the right, it's a good decision to drive on the right too.

NASH EQUILIBRIUM

The outcome when each individual plays his or her best response to the strategies chosen by everyone else.

Named after John Nash, a mathematician and economist.

Multiple Nash equilibria may arise even in simple economic problems, such as the choice of crops by Bala and Anil. Consider a situation different from the invisible hand game we played in Figure 4.2. In the new payoff matrix, if the two farmers produce the same crop, there is now such a large fall in price that it is better for each to specialise, even if in the crop they are less suited to grow.

Situations with two Nash equilibria prompt us to ask two questions:

- Which equilibrium would we expect to observe in the world?
- Is there a conflict of interest because one equilibrium is preferable to some players, but not to others?

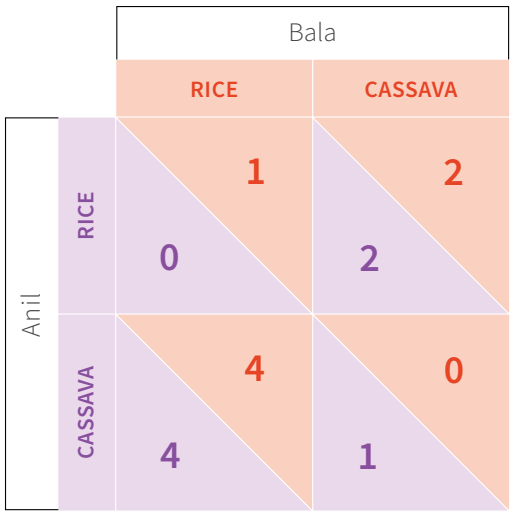


Figure 4.14 *A division of labour problem in which the invisible hand might not work: More than one Nash equilibrium.*

Whether you drive on the right or the left is not a matter of conflict in itself, as long as everyone you are driving towards has made the same decision as you. We can't say that driving on the left is better than driving on the right.

In the division of labour game faced by Anil and Bala, unlike the driving game, it is clear that the Nash equilibrium with Anil choosing *Cassava* and Bala *Rice* (where they specialise in their better crop) is preferred to the other Nash equilibrium by both farmers.

Could we say, then, that we would expect to see Anil and Bala engaged in the "correct" division of labour? Not necessarily. Remember, we are assuming that they take their decisions independently, without coordinating with each other. Imagine that Bala's father had been especially good at growing cassava (unlike his son) and so the land (although

better suited to rice) remained dedicated to cassava. In response to this, Anil knows that *Rice* is his best response to Bala's *Cassava*, and so would have then chosen to grow rice. Bala would have no incentive to switch to what he is good at: growing rice.

The example makes an important point: If there is more than one Nash equilibrium, and if people choose their actions independently, then an economy can get "stuck" at a Nash equilibrium in which *all are worse off than they would be at the other equilibrium*.

GREAT ECONOMISTS

JOHN NASH

John Nash (1928-2015) completed his doctoral thesis at Princeton University at the age of 21. It was only 27 pages long, yet it gave game theory— then a little-known mathematical language—an application in economics. He provided an answer to the question: When people interact strategically, what would you expect them to do? His answer, now known as a *Nash equilibrium*, is that we would expect to see outcomes of games with the property that no player would prefer to play differently, given the actions of each of the other players.

Nash would go on to share a Nobel prize for his work. Roger Myerson, an economist who also won a Nobel prize, described the Nash equilibrium as "one of the most important contributions in the history of economic thought."

Nash originally wanted to be an electrical engineer like his father, and studied mathematics as an undergraduate at Carnegie Tech (now Carnegie-Mellon University). An elective course in International Economics began his interest in how bargaining problems were resolved, which eventually led to his breakthrough.

For much of his life Nash suffered from mental illness that required hospitalisation. He experienced hallucinations caused by schizophrenia that began in 1959, though after what he described as "25 years of partially deluded thinking" he continued his teaching and research at Princeton. The story of his insights and illness are told in the book (made into a film starring Russell Crowe) *A Beautiful Mind*.

Resolving conflict

Although Bala and Anil prefer the same Nash equilibrium (in which they specialise in the right crops), when there is more than one equilibrium, there may be a conflict of interest over which one should occur.

To see this we take a different example. Consider the case of Astrid and Bettina, two software engineers who are working on a project for which they will be paid. Their first decision is whether the code should be written in Java or C++ (imagine that either programming language is equally suitable, and that parts of the application can be written in each language). They each have to choose one or the other, and Astrid wants to write in Java because she is better at writing Java code. While this is a joint project with Bettina, her pay will be based (in part) on how many lines of code were written by her. Unfortunately Bettina prefers C++, for just the same reason. So the two strategies are called *Java* and *C++*.

Their interaction is described in Figure 4.15a, and their payoffs are in Figure 4.15b.

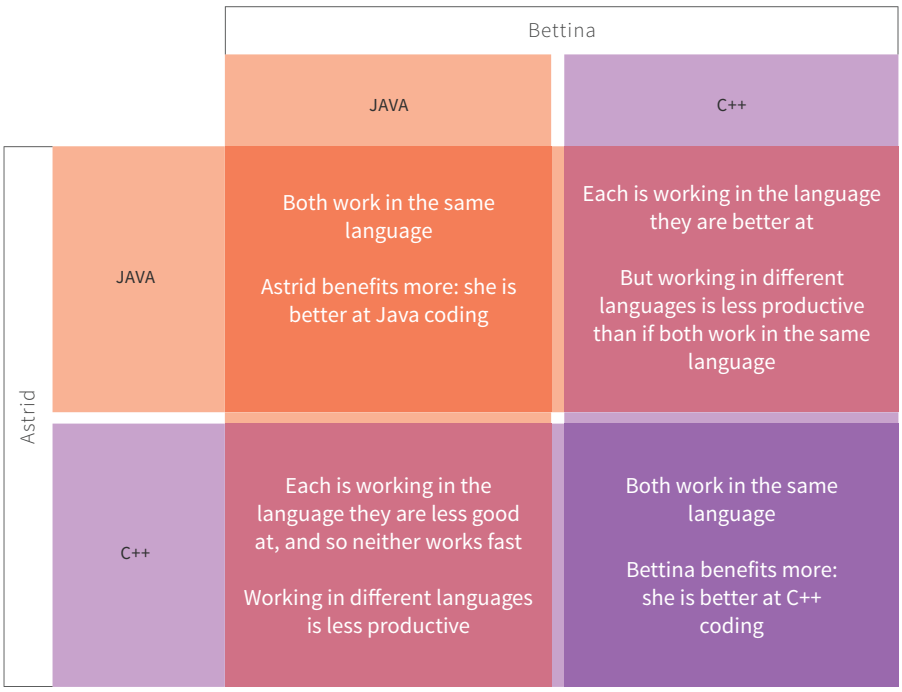


Figure 4.15a *Conflict over the choice of programming language when there is more than one Nash equilibrium.*

From Figure 4.15b, you can work out three things:

- They both do better if they work in the same language.
- Astrid does better if that language is *Java*, while the reverse is true for Bettina.
- Their total payoff is higher if they choose C++.

How would we predict the outcome of this game?

Think through each of the possible outcomes to see if some can be eliminated. If Astrid chose *C++*, then Bettina surely would not choose *Java*: they'd both get zero because they'd be working in different languages *and* in the ones that they were least good at. We can eliminate the (*0*, *0*) outcome. Similar reasoning will let you eliminate the outcome in which Bettina chooses *C++* while Astrid chooses *Java* (*2*, *2*). They do a little better here than in the opposite case, because at least each is working in the language with which they are familiar.

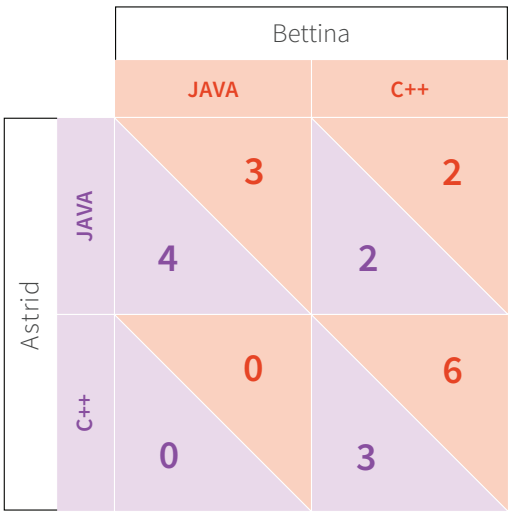


Figure 4.15b *Payoffs in the conflict over the choice of programming language when there is more than one Nash equilibrium. The payoffs show the pay in thousands of dollars to complete the project.*

It is also true that Astrid's best response to Bettina choosing *Java* is to choose *Java* too, and vice versa (*4*, *3*). And Astrid's best response to Bettina's choice of *C++* is *C++*, and vice versa (*3*, *6*).

So the two outcomes in which they work in different programming languages can be ruled out. This is because they are not Nash equilibria, that is, they are *not mutual best responses*.

But what of the two Nash equilibria in which the two work in the same language? Astrid obviously prefers that they both play *Java* while Bettina prefers that they both play *C++*. With the information we have about how the two might interact, we can't yet predict what would happen. While we have considered here cases in which there are two Nash equilibria, in many interactions there will be just one. It need not be a dominant strategy equilibrium (as in the prisoners' dilemma) but it is still a reasonable prediction of what we should observe in the interaction, because each of the two are doing the best they can given what the other is doing. Discuss 4.9 gives some examples of the type of information that would help to clarify what we would observe.

DISCUSS 4.10: CONFLICT BETWEEN ASTRID AND BETTINA

What is the likely result of this game if:

1. Astrid can choose which language she will use first, and commit to it (just as the Proposer in the ultimatum game commits to an offer, to which the Responder then replies by accepting or rejecting)?
2. The two can make an agreement, including which language they use, and the size of a cash transfer from one to the other?
3. They have been working together for many years, and in the past they used Java on joint projects?

DISCUSS 4.11: CONFLICT IN BUSINESS

In the 1990s, Microsoft battled Netscape over market share for their web browsers, called Internet Explorer and Navigator. In the 2000s, Google and Yahoo fought over which company's search engine would be more popular. In the entertainment industry a battle called the "format wars" played out between Blu-Ray and HD-DVD.

Use one of these examples to analyse whether there are multiple equilibria and, if so, why one equilibrium might emerge in preference to the others.

4.12 CONCLUSION

The irrigation system in Valencia, Spain, seemed destined for overuse and decline, a tragedy of the commons waiting to happen, but as we saw in the introduction, Garrett Hardin's dismal drama never played out there. The *Tribunal de las Aguas* regulated water use and has preserved the resource for hundreds of years.

Spanish farmers brought similar institutions to the new world, where communities still sustain irrigation channels and regulate sustainable water use in the state of New Mexico in the US. Similar institutions are found across the world throughout human history, from forests in the Italian Alps in the 13th century that were successfully managed by community contractual systems, to the recovery of whale stocks in recent times based on voluntary international agreements.

Even present-day global environmental problems have sometimes been tackled effectively. The *Montreal Protocol* to phase out and eventually ban the chlorofluorocarbons (CFCs) that threatened to destroy the ozone layer (which protects us against harmful ultraviolet radiation) has been remarkably successful.

CONCEPTS INTRODUCED IN UNIT 4

Before you move on, review these definitions:

- Game
- Best response
- Dominant strategy equilibrium
- Social dilemma
- Altruism
- Reciprocity
- Inequality aversion
- Nash equilibrium
- Public goods
- Prisoners' dilemma

Institutions are not always able to solve social dilemmas. The success of the Montreal Protocol in limiting CFCs contrasts with the relative failure of the Kyoto Protocol for reducing carbon emissions responsible for global warming. The reasons are likely to be scientific and political, as well as economic: for example, the alternative technologies to CFCs were well-developed and the benefits relative to costs for large industrial countries, such as the US, were much clearer and larger than in the case of greenhouse gas emissions.

As the *Stern Review* made clear, the problem of climate change is far from a solution. Economics in general, and especially game theory, can help us understand some of the obstacles to implementing a solution.

Remember: economics is the study of how people interact with nature and with one another in producing our livelihoods. Part of the problem with how we now relate to nature (for example, causing climate change) can be traced to how we relate to each other (in this case, a failure to implement adequate global regulation of greenhouse gases). Game theory applied to the study of social interactions provides some clues as to why this is occurring.

Recall that Anil and Bala faced a prisoners' dilemma resulting in both using the pesticide *Terminator* because they were:

- *Self-interested* and so did not internalise the harm that their decisions would inflict
- *Not subject to any form of peer punishment* for the harm that *Terminator* did to the payoffs of the other person
- *Unable to make an agreement* simply banning the use of *Terminator*.

These assumptions capture many of the difficult realities facing people in every country in the world who are seeking to control climate change.

Think of the problem of climate change as a game between two "countries" called China and United States, considered as if each were a single individual. Each country has two possible strategies for addressing global carbon emissions: *Restrict* and *BAU* (the Stern report's *business as usual* scenario):

- *Restrict* could be implemented in a variety of ways that we consider in more depth in Unit 18, including limits on the quantity of fossil fuels that can be used, or policies to raise the price of fossil fuels to induce firms and others to economise on their use.
- *BAU* means making no change in existing policies.

The four possible outcomes are indicated in Figure 4.16:

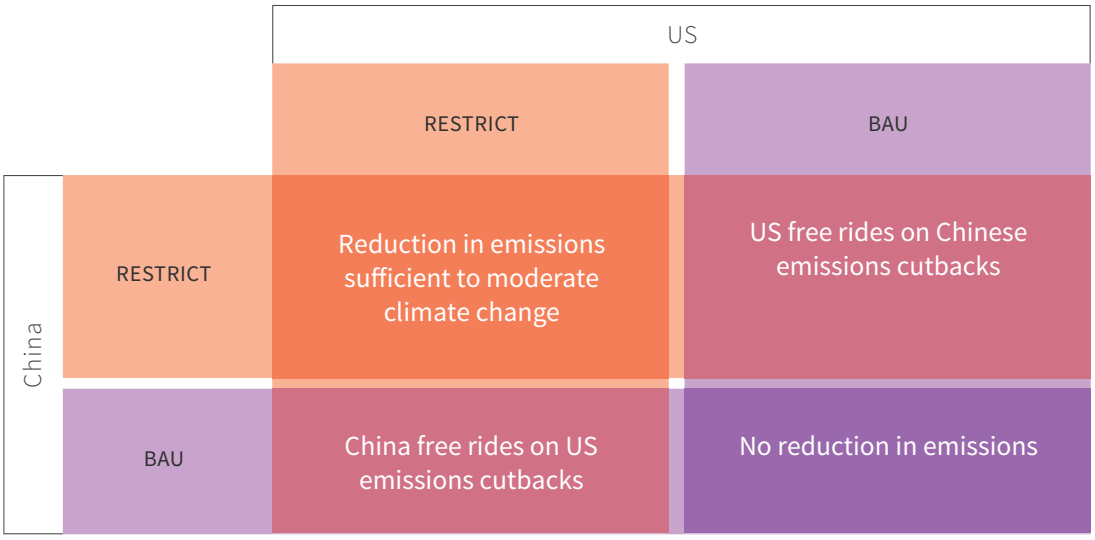


Figure 4.16 *Climate change policy as a prisoners' dilemma.*

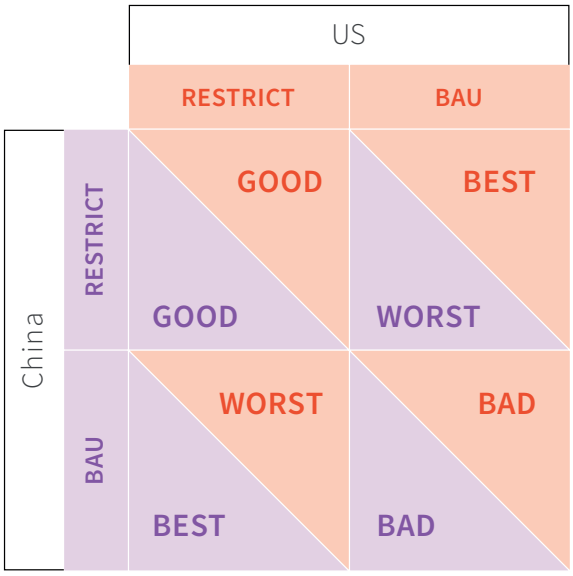
In Figure 4.17 we give hypothetical payoffs for the two countries on a scale from *best*, through *good* and *bad*, to *worst*. This is called an ordinal scale (because all that matters is the order: that *best* is better than *good*, for example, not by how much it is better). You can see them in the first panel of Figure 4.17.

An ordinal scale does not allow us to add up the payoffs for China and the US (for example, we do not know if *good* for both is better in total than *best* for one and *worst* for the other). Ordinal measures of preferences are commonly used in economics and, in this case, they give us enough information to surmise how each country will play the game:

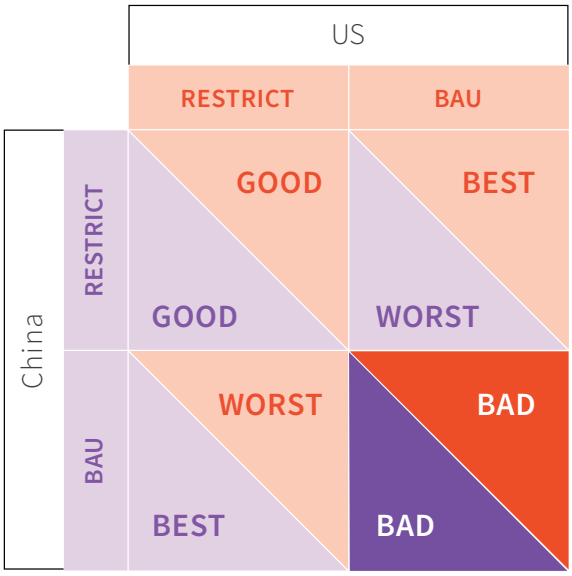
- *Self-interest makes BAU a Nash equilibrium*: If one nation restricted its emissions the climate change problem would be sufficiently lessened so that the other country would prefer not to bear the costs of emissions limitation. If this is the

case, then a little study of the payoffs will convince you that China and the US are in a prisoners' dilemma. BAU is the dominant strategy for both, so the dominant strategy equilibrium is to continue with existing policies. Check this on the second panel of Figure 4.17.

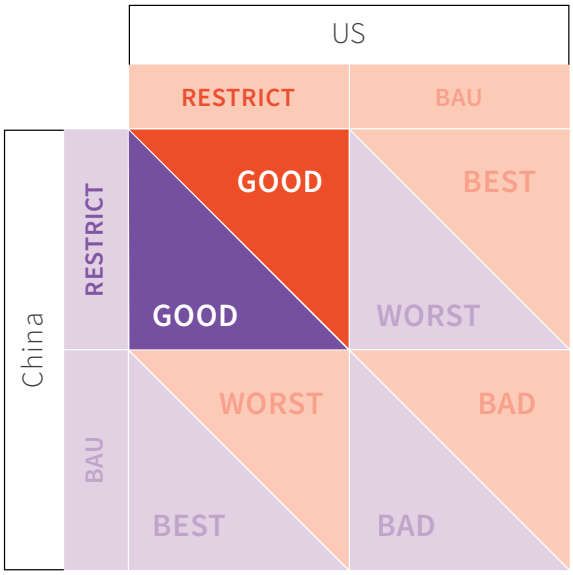
- *International agreements may eliminate BAU*: A formally binding treaty might address this problem by simply eliminating the BAU strategy. This transforms the social interaction between China and the US as shown in the third panel of Figure 4.17. If this happens it's not a game at all—neither the US nor China has any choice among strategies.



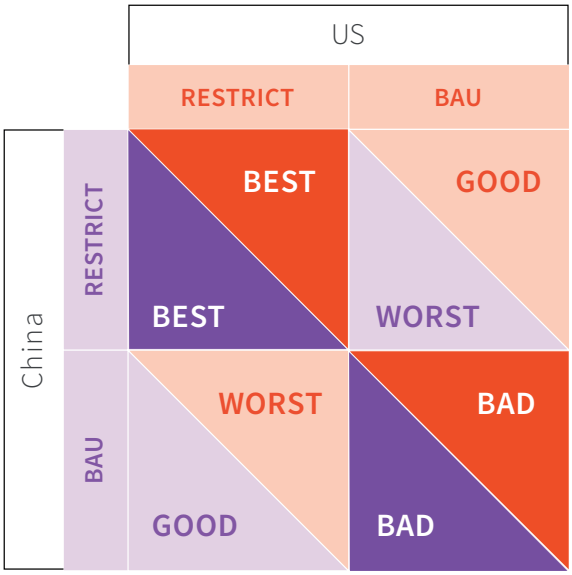
(1) The climate change game's payoffs



(2) Business as usual as a dominant strategy equilibrium



(3) The effect of a treaty



(4) Self-interest with inequality aversion and reciprocity

Figure 4.17 *Payoffs for climate change policy as a prisoners' dilemma.*

But, if American and Chinese people cared only about the wellbeing of their fellow citizens, it is unlikely that such treaties would ever be negotiated, and if they were negotiated it would be unlikely that they were observed. The problem of gaining support for such a treaty and implementing it were it signed would perhaps be surmountable if the Chinese and Americans were not entirely self-interested. Suppose that in both countries people were also both *inequality averse* and *reciprocal*. How would this change the payoff matrix?

- *The effect of inequality aversion and reciprocity*: If they were inequality averse, then the previous worst outcome—a situation in which their nation bears the costs of emissions limitation but the other does not—is even worse than before. By itself, this will not alter how the game is played: *BAU* remains the dominant strategy equilibrium. But more importantly, their inequality aversion would also lead them to value less highly the situation in which the inequality is reversed. If they were also motivated by reciprocity they might be willing to bear the costs of restricting emissions—but only on the condition that the other country did the same. This creates the possibility that there are two Nash equilibria in the game.
- *Two Nash equilibria when people are also reciprocal and inequality averse*: The combination of reciprocity and inequality aversion has made the asymmetric outcomes less attractive to the countries, so that both countries restricting emissions is now a Nash equilibrium. You can confirm, too, that *BAU* remains a Nash equilibrium, but it is no longer a dominant strategy equilibrium. So there are two Nash equilibria: the payoff matrix looks like the final panel in Figure 4.17.

The physical, economic and political complexities of the problem of climate change cannot, of course, be fully represented in a simple two-person, one-shot game.

But the example of carbon emission restriction illustrates the ability of game theory to clarify some of the possible approaches to bringing about desirable outcomes when people or nations interact, and also to understanding why the desirable outcomes sometimes do not occur.

DISCUSS 4.12: NASH EQUILIBRIA AND CLIMATE CHANGE

1. Describe the changes in preferences or in some other aspect of the problem that would convert the game to one in which (like the invisible hand game) both adopting *Restrict* is a dominant strategy equilibrium.
2. Can you think of any circumstances under which the asymmetric outcomes (one country plays *Restrict*, the other adopts *BAU*) would be the Nash equilibrium (*Hint*: think of the game of "chicken")?

Key points in Unit 4

Game theory

Many social interactions—such as competition between large firms, relationships between employers and employees, and how human economic activity affects climate change—can be studied using game theory.

The invisible hand

Under some rules of the game, the pursuit of self-interest results in mutually beneficial outcomes.

Social dilemmas

The public goods game and the prisoners' dilemma are social dilemmas.

Unfortunate outcomes of social dilemmas

In social dilemmas the outcome is worse for all participants than an alternative feasible outcome that could result if the players were motivated by social preferences, or subject to peer punishment, or could make binding agreements about the strategies they choose.

Social preferences

Behavioural experiments show that many people are motivated by social preferences such as reciprocity, inequality aversion, and altruism, as well as by self-interest.

Preference for fairness

Ultimatum game experiments show that people would often prefer to forego substantial personal gains rather than to receive an unfair share of the total mutual gains.

The role of institutions

Economic and political institutions—the rules of the game—matter for attaining socially valued outcomes when people face social dilemmas such as global climate change.

4.13 EINSTEIN

When will an offer in the ultimatum game be accepted?

Suppose $100 is to be split, and there is a fairness norm of 50-50. When the proposal is $50 or above, ($y \geq 50$), the Responder feels positively disposed towards the Proposer and would naturally accept the proposal, as rejecting it would hurt both herself and also someone she appreciates because they conform to, or were even more generous than, the social norm. But if the offer is below $50, ($y < 50$) then she feels the 50-50 norm is not being respected, and she may want to punish the Proposer for this breach. If she does reject the offer, this will come at a cost to her: rejection means both leave with nothing.

To make the situation concrete, let us suppose her anger at the breach of the social norm depends on the size of the breach: if the Proposer offers nothing she will be furious, but she's more likely to be puzzled than angry at an offer of $49.50 rather than the $50 offer she might have expected had the norm been followed. So how much satisfaction she would derive from punishing a Proposer's low offer depends on two things. The first is R, a number that indicates how strong is her private reciprocity motive: if R is a large number then she cares a lot about whether the Proposer is acting generously and fairly or not; if $R = 0$ she not at all reciprocal. So the satisfaction at rejecting a low offer is $R(50 - y)$. The gain from accepting the offer is the offer itself, or y.

The decision to accept or reject just depends on which of these two quantities is larger. We can write this as "reject an offer if $y < R(50 - y)$". This equation says: she will reject an offer of less than $50 according to how much lower than $50 the offer is (as measured by $(50 - y)$ multiplied by her private attitude to reciprocity, R).

To calculate her minimum acceptable offer we can rearrange this rejection equation like this:

$$\begin{aligned} y &< R(50 - y) \\ y &< 50R - Ry \\ y + Ry &< 50R \\ y(1 + R) &< 50R \\ y &< \frac{50R}{(1 + R)} \end{aligned}$$

If $R = 1$, then $y < 25$ and she will reject any offer less than $25. This makes intuitive sense if her attitude to reciprocity is in line with the 50-50 social norm: if she rejects the offer of $25, she loses $25 and splits the difference 50-50 with the Proposer between rejecting the offer and an offer of $50, which is the social norm.

If $R = 2$, then $y < 33.33$ and she will reject any offer less than \$33.33. A value of R above one means she places more weight on reciprocity than the social norm and the offer has to be higher for her not to reject it. Similarly for a value of $R < 1$. For $R = 0.5$, for example, $y < 16.67$ and offers below \$16.67 will be rejected.

DISCUSS 4.12: ACCEPTABLE OFFERS

1. Draw the game tree for the game proposed above.
2. How might the Minimum Acceptable Offer depend on the method by which the Proposer acquired the \$100 (for example: did she find it on the street, win it in the lottery, receive it as an inheritance, and so on)?
3. Suppose that the fairness norm in this society is 50-50. Can you imagine anyone offering more than 50% in such a society?
4. If so, why?

4.14 READ MORE

Bibliography

1. Aesop. 2013. *Belling the Cat*. United States: Picture Window Books.
2. Bewley, Truman. 2007. 'Fairness, Reciprocity, and Wage Rigidity.' *Behavioral Economics and Its Applications*, 157–88.
3. Bowles, Samuel, and Herbert Gintis. 2011. *A Cooperative Species: Human Reciprocity and Its Evolution*. Princeton, NJ: Princeton University Press.
4. Brosnan, Sarah F., and Frans B. M. de Waal. 2003. 'Monkeys Reject Unequal Pay.' *Nature* 425 (6955): 297–99.
5. Camerer, Colin, and Ernst Fehr. 2004. 'Measuring Social Norms and Preferences Using Experimental Games: A Guide for Social Scientists.' In *Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies*, edited by Joseph Henrich, Robert Boyd, Samuel Bowles, Colin Camerer, and Herbert Gintis. Oxford: Oxford University Press.
6. Edgeworth, Francis Ysidro. (1881) 2003. *Mathematical Psychics and Further Papers on Political Economy*. Oxford: Oxford University Press.
7. Falk, Armin, and James J. Heckman. 2009. 'Lab Experiments Are a Major Source of Knowledge in the Social Sciences.' *Science* 326 (5952): 535–38.

8. Fehr, Ernst, and Urs Fischbacher. 2003. 'The Nature of Human Altruism.' *Nature* 425 (6960): 785–91.
9. Fischbacher, Urs, Christina M. Fong, and Ernst Fehr. 2009. 'Fairness, Errors and the Power of Competition.' *Journal of Economic Behavior & Organization* 72 (1): 527–45.
10. Hardin, Garrett. 1968. 'The Tragedy of the Commons.' *Science* 162 (3859): 1243–48.
11. Henrich, Joseph, Robert Boyd, Samuel Bowles, Colin Camerer, and Herbert Gintis, eds. 2004. *Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies*. Oxford: Oxford University Press.
12. Henrich, Joseph, Richard McElreath, Abigail Barr, Jean Ensminger, Clark Barrett, Alexander Bolyanatz, Juan Camilo Cardenas, et al. 2006. 'Costly Punishment Across Human Societies.' *Science* 312 (5781): 1767–70.
13. Herrmann, Benedikt, Christian Thoni, and Simon Gachter. 2008. 'Antisocial Punishment Across Societies.' *Science* 319 (5868): 1362–67.
14. IPCC. 2014. *Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Geneva, Switzerland: IPCC.
15. Levitt, Steven D., and John A. List. 2007. 'What Do Laboratory Experiments Measuring Social Preferences Reveal About the Real World?' *Journal of Economic Perspectives* 21 (2): 153–74.
16. Mencken, H. L. (1916) 2006. *A Little Book in C Major*. New York, NY: Kessinger Publishing.
17. Nasar, Sylvia. 2011. *A Beautiful Mind: The Life of Mathematical Genius and Novel Laureate John Nash*. New York, NY: Simon & Schuster.
18. Ostrom, Elinor. 2000. 'Collective Action and the Evolution of Social Norms.' *Journal of Economic Perspectives* 14 (3): 137–58.
19. Ostrom, Elinor. 2008. 'The Challenge of Common-Pool Resources.' *Environment: Science and Policy for Sustainable Development* 50 (4): 8–21.
20. Ostrom, Elinor, James Walker, and Roy Gardner. 1992. 'Covenants With and Without a Sword: Self-Governance Is Possible.' *The American Political Science Review* 86 (2).
21. Shiller, Robert J., and George A. Akerlof. 2009. *Animal Spirits: How Human Psychology Drives the Economy, and Why It Matters for Global Capitalism*. 9th ed. Princeton, NJ: Princeton University Press.
22. Stern, Nicholas. 2007. *The Economics of Climate Change: The Stern Review*. Cambridge: Cambridge University Press.
23. Thaler, Richard. 2015. *Misbehaving: The Making of Behavioral Economics*. New York, NY: W. W. Norton.

Beta January 2016 version

5

PROPERTY AND POWER: MUTUAL GAINS AND CONFLICT

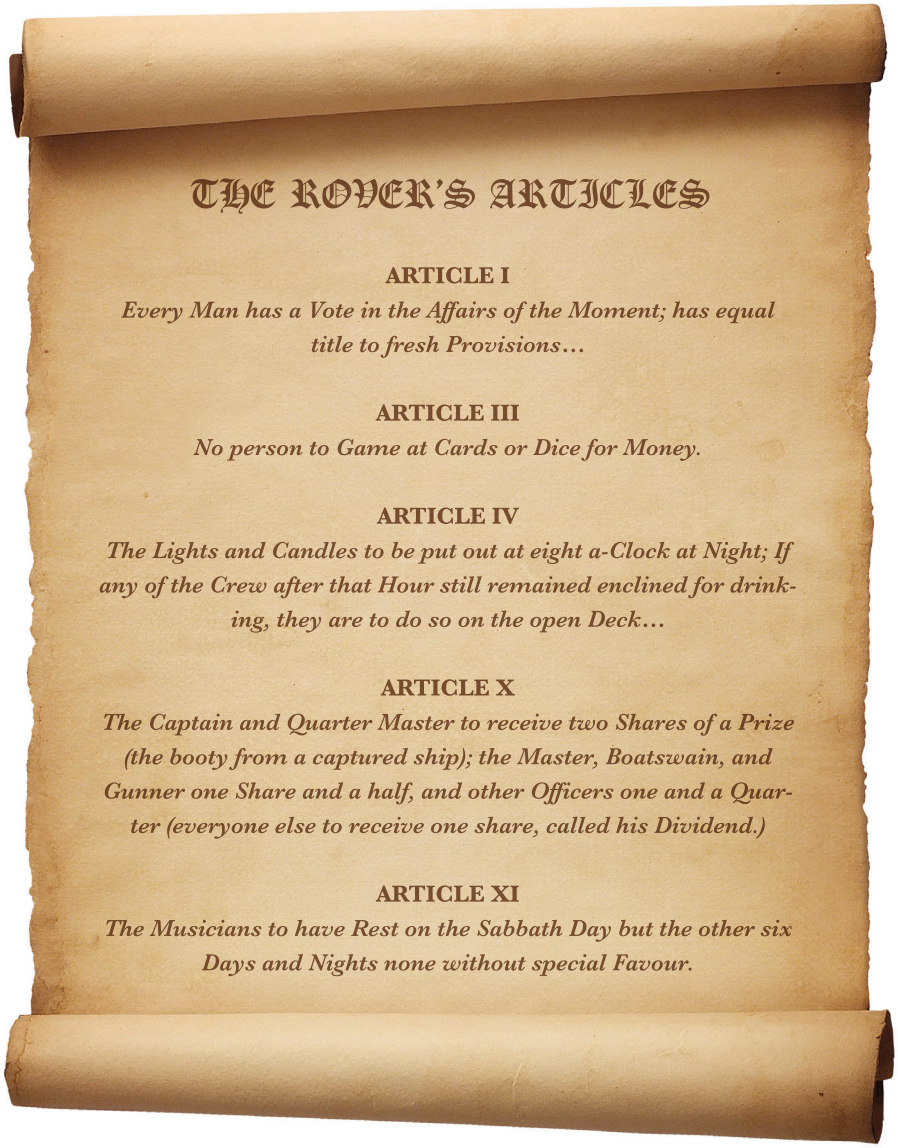
cc by muffinn, Flickr

HOW INSTITUTIONS INFLUENCE THE BALANCE OF POWER IN INTERACTIONS AMONG ECONOMIC ACTORS, AND HOW THIS AFFECTS THE FAIRNESS AND EFFICIENCY OF THE ALLOCATIONS THAT RESULT

- Technology, biology, economic institutions and people's preferences all matter as determinants of economic outcomes
- Interactions between economic actors can result in mutual gains, and also in conflicts over how the gains are distributed
- *Power* is the ability to do and get the things we want in opposition to others
- Institutions influence the power and other bargaining advantages of actors
- Outcomes may be judged according to their efficiency and their fairness
- Economics can clarify ways of applying the criteria of efficiency and fairness to evaluate economic institutions and outcomes

Institute for New Economic Thinking
UCL SciencesPo Azim Premji University
FRIENDS PROVIDENT Foundation

Perhaps one of your distant ancestors considered the best way to get money was by shipping out with a pirate like Blackbeard or Captain Kidd. If he had settled on Captain Bartholomew Roberts' pirate ship *The Rover*, he and the other members of the crew would have been required to consent to the ship's written constitution. This document (called *The Rover's Articles*) guaranteed, among other things, that:



THE ROVER'S ARTICLES

ARTICLE I

Every Man has a Vote in the Affairs of the Moment; has equal title to fresh Provisions…

ARTICLE III

No person to Game at Cards or Dice for Money.

ARTICLE IV

The Lights and Candles to be put out at eight a-Clock at Night; If any of the Crew after that Hour still remained enclined for drinking, they are to do so on the open Deck…

ARTICLE X

The Captain and Quarter Master to receive two Shares of a Prize (the booty from a captured ship); the Master, Boatswain, and Gunner one Share and a half, and other Officers one and a Quarter (everyone else to receive one share, called his Dividend.)

ARTICLE XI

The Musicians to have Rest on the Sabbath Day but the other six Days and Nights none without special Favour.

Source: *Leeson, Peter T. 2007. 'An–arrgh–chy: The Law and Economics of Pirate Organization.' Journal of Political Economy 115 (6): 1049–94.*

The *Rover* and its *Articles* were not unusual. During the heyday of European piracy in the late 17th and early 18th centuries most pirate ships had written constitutions that guaranteed even more powers to the crew members. Their captains were democratically elected, "the Rank of Captain being obtained by the Suffrage of the

Majority". Many captains were also voted out, at least one for cowardice in battle. Crews also elected one of their number as the quartermaster who, when the ship was not in a battle, could countermand the captain's orders.

If your ancestor had served as a lookout and had been the first to spot a ship that was later taken as a prize, he would have received as a reward "the best Pair of Pistols on board, over and above his Dividend". Were he to have been seriously wounded in battle, the articles guaranteed him compensation for the injury (more for the loss of a right arm or leg than for the left). He would have worked as part of a multiracial, multi-ethnic crew of which probably about a quarter were of African origin, and the rest primarily of European descent, including Americans.

The result was that a pirate crew was often a close-knit group. A contemporary observer lamented that the pirates were "wickedly united, and articled together". Sailors of captured merchant ships often happily joined the "roguish Commonwealth" of their pirate captors.

Nowhere else in the world during the late 17th and early 18th century did ordinary workers have the right to vote, or to compensation for occupational injuries, or to the protection of the kinds of checks on arbitrary authority that were taken for granted on *The Rover*.

Nor could workers in British textile mills and other industrial establishments claim such a large share of income. The prize-sharing system described in The Rover's articles, if faithfully implemented, would have resulted in a Gini coefficient for the dividend of 0.06, far more equal even than our famously equal hunter-gatherer ancestors.

In contrast, when the Royal Navy's ships *Favourite* and *Active* captured the Spanish treasure ship *La Hermione*, the division of the spoils among the captain, officers and crew of the two British men-of-war ships resulted in a Gini coefficient averaging 0.61 for the two ships: about the same as the Gini coefficient for income in some of the most unequal countries in the world today (shown in Figure 1.16). By the standards of the day, pirates were unusually democratic and fair-minded in their dealings with each other.

Another unhappy commentator remarked: "These Men whom we term... the Scandal of human Nature, who were abandoned to all Vice... were strictly just among themselves." If they were Responders in the ultimatum game, by this description they would have rejected any offer less than half of the pie!

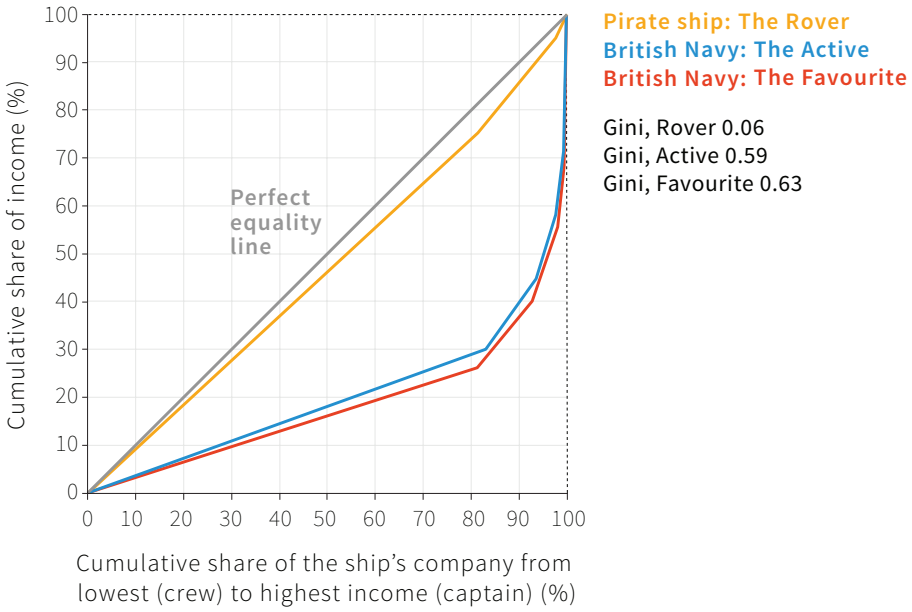


Figure 5.1 *Inequality in the division of the spoils: Pirates and the Royal Navy.*

5.1 INSTITUTIONS: THE RULES OF THE GAME

The Rover's articles were part of the pirate institutions that determined who did what aboard ship—whether your ancestor would serve as a lookout or as a helmsman, for example—and who would get what as a result of what each did, for example the size of his dividend compared to that of the gunner. Other aspects of their institutions were the unwritten informal rules of appropriate behaviour that the pirates followed by custom, or to avoid condemnation by their crewmates.

Using the terminology of game theory introduced in the previous unit, we could say that *The Rover*'s articles were the rules of the game, much as the rules of the ultimatum game specify who can do what, when they can do it, and how what each player does determines each player's payoff. The institutions provided both the *constraints* (no drinking after 8pm unless on deck) and the *incentives* (the best pair of pistols for the lookout who spotted a ship that was later taken). In this unit we use the terms *institutions* and *rules of the game* interchangeably.

INSTITUTIONS

Institutions are written and unwritten rules that govern:

- What people do when they interact in a joint project
- The distribution of the products of their joint effort

Experiments showed us in section 4.6 that the rules of the game affect:

- How the game is played
- The size of the total available to those participating
- How this total is divided

For example, the rules (institutions) of the ultimatum game determine who gets to be the Proposer; how much money the Proposer has when the game starts; and the fact that the Responder can refuse any offer, resulting in no payoffs for either player. In the standard ultimatum game, with a Proposer and one Responder, recall that the total to be divided may be zero for both players if the Responder refuses the Proposer's offer. Or, if the Proposer's offer is accepted, the Responder's share is the amount that the Proposer offers to share, while the Proposer gets what remains.

We also saw in section 4.10 that *changing the rules* changes the outcome: if there are two Responders in the ultimatum game rather than just one, the Proposer knows that at least one of the Responders is likely to accept a low offer. Each Responder knows this too. And because they cannot be sure that their rejecting a low offer will result in the proposer being punished (the other Responder may accept). Responders tend to accept low offers, which they would have rejected as unfair had they been the sole Responder. This means the Proposer has more bargaining power, and gets a larger payoff as a result. We will discover in Unit 7 and Unit 8 that, when people must purchase goods from a single business organisation, they have less bargaining power than when there are many sellers.

In the ultimatum game (and in the economy) the division of the payoffs depends on what is called *bargaining power*. A party's bargaining power is the extent of their advantage when dividing the pie. The Proposer in the ultimatum game has the fortunate position of making the offer, and so is likely to get at least half of the pie, while the most the Responder is likely to get is half. Being the Proposer means having more bargaining power, but the power is limited. The Proposer's bargaining power is limited by the need to get the Responder to accept the offer.

The Proposer could have even more than this take-it-or-leave-it power: the rules might allow a Proposer simply to divide up the pie in any way, without any role for the Responder other than to take whatever he gets (if anything). In this case the Proposer has all of the bargaining power and the Responder none. There is an experimental game like this, and it is called (you guessed it) the *dictator game*.

The past, and even the present, provide many examples of economic institutions that are like the dictator game, in which there is no option to say no. Examples include today's remaining political dictatorships such as North Korea, and slavery as it existed in the US prior to the end of the American Civil War in 1864. Criminal organisations involved in drugs and human trafficking would be another modern example.

Unit 2 showed that the pay people receive for their work depends on the rules of the game as well as technology. Remember from Unit 1 that productivity of labour started to increase in Britain around the middle of the 17th century. But it was not until the middle of the 19th century that a combination of shifts in the supply and demand for labour and new institutions, such as trade unions and the right to vote for workers, gave wage earners the bargaining power to raise wages substantially.

5.2 EVALUATING INSTITUTIONS AND OUTCOMES: PARETO EFFICIENCY

Whether it is fishermen seeking to make a living while not depleting the stocks of fish, or farmers maintaining the channels of an irrigation system, or two people dividing up a pie, we typically have social norms about what ought to happen. We describe these situations in two ways: what actually happens, and an evaluation of whether it is good by some standard. The first involves facts; the second involves values.

We call the outcome of an economic interaction an *allocation*.

For the ultimatum game the allocation describes the proposed division of the pie by the Proposer, whether it was rejected or accepted, and the resulting payoffs to the two players.

It is often important to go beyond a description of the allocation and to evaluate the outcome: how good is it? An allocation can be evaluated from two standpoints: *efficiency* and *fairness*.

ALLOCATION

An *allocation* is:

- A description of who does what
- Plus the consequences of their actions...
- ... including who gets what

PARETO EFFICIENCY

Named for Vilfredo Pareto, an Italian economist and sociologist, this describes an allocation with the property that there is no alternative allocation in which at least one party could be better and nobody worse off.

For an engineer, efficiency means the most sensible way to go about achieving something, for example, producing electricity at the least cost or making the most of the use of some scarce resource. This is not what economists mean by the term. For any allocation, an economist interested in efficiency asks whether there is some other allocation in which all of the parties could be better off (each party prefers the allocation), or at least one of them could be better off and none worse off. If there is no allocation for which this is possible, we say the current allocation is *Pareto efficient*.

If something is not Pareto efficient, naturally we call it *Pareto inefficient*. It's clear that a Pareto inefficient allocation is not one that we would favour: there is, by definition, an alternative in which at least one party would be better off, and no one would be worse off.

GREAT ECONOMISTS

VILFREDO PARETO

Vilfredo Pareto (1848-1923), an Italian economist and sociologist, earned a degree in engineering for his research on the concept of equilibrium in physics. He is mostly remembered for the concept of efficiency that bears his name. He wanted economics and sociology to be fact-based sciences, similar to the physical sciences that he had studied at first.

His empirical investigations led him to question the idea that the distribution of wealth resembles the familiar bell curve with a few rich and a few poor in the tails of the distribution, and a large middle-income class. In its place he proposed what came to be called *Pareto's law* according to which, across the ages and differing types of economy, there were very few rich people, and a lot of poor people.

His *80-20 rule*—derived from Pareto's law—asserted that the richest 20% of a population typically held 80% of the wealth. Were he living in the US in 2015, he would have to revise that to 90% of the wealth held by the richest 20%, suggesting that his law might not be as universal as he had thought.

In Pareto's view, the economic game was played for high stakes, with big winners and losers. Not surprisingly, then, he urged economists to study conflicts over the division of goods, and he thought the time and resources devoted to these conflicts were part of what economics should be about. In his most famous book, the *Manual of Political Economy*, he wrote that:

"[T]he efforts of men are utilised in two different ways: they are directed to the production or transformation of economic goods, or else to the appropriation of goods produced by others."
Vilfredo Pareto, *A Manual of Political Economy* (1906)

His lifelong interest was political and economic inequality, which he combined with a growing hostility towards socialism, trade unions, and government interventions in the economy, eventually leading before his death to some sympathies for the rising Italian fascist movement. He was a pioneer in economic theory, the combination of economics with insights from political science and sociology, and in the empirical estimation of economic quantities.

The difference between Pareto efficient and Pareto inefficient allocations is clear in the prisoners' dilemma game played by Anil and Bala in Unit 4, shown in Figure 5.2. To determine if an allocation is Pareto efficient we draw a rectangle with a corner at the point in question, say the point (*I, T*) at which Anil plays *IPC* and Bala plays *Terminator*. The rectangle covers the area to the north-east of the point. We ask: is there any feasible outcome in the rectangle? If there is no feasible outcome in this space, then no win-win change from the point (*I, T*) is possible, and the allocation is Pareto efficient. Figure 5.2 shows you how to check Pareto efficiency for each of the four possible allocations.

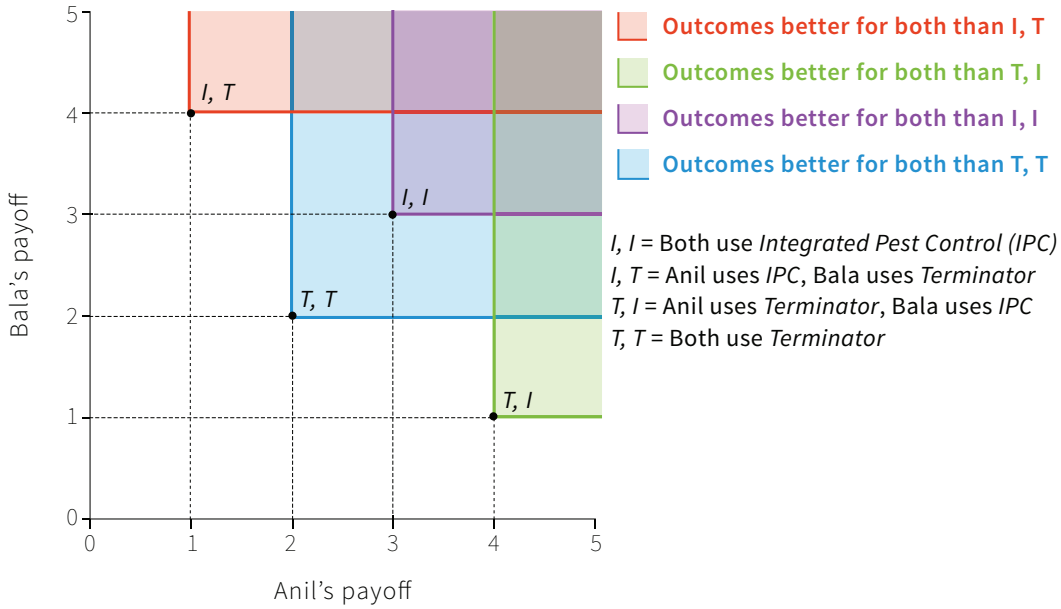


Figure 5.2. *Pareto efficient allocations. All of the allocations except mutual pesticide (T,T) are Pareto efficient.*

So Anil playing IPC and Bala playing Terminator is Pareto efficient. Anil may think this is unfair. Even Bala may think it is unfair. Pareto efficiency *has nothing to do with fairness.*

The same is true of the situation in which Anil uses *Terminator* and Bala chooses *IPC* (*T, I*). And both doing *IPC* (I, I) is also Pareto efficient. The only point that is not Pareto efficient is when both use *Terminator* (*T, T*) because both could be better off if

they both used IPC: the point (I, I) is in the shaded rectangle whose corner is at (T, T). Some people might disapprove of outcomes in which free riding occurs, though both outcomes are Pareto efficient.

There are many Pareto efficient allocations that we would not evaluate favourably. In section 4.4 we saw in Figure 4.3 that any split of Anil's lottery winnings (including giving Bala nothing) is Pareto efficient (to see this, choose any point on the boundary of the feasible set of outcomes, and draw the rectangle with its corner at that point, just as we have done in Figure 5.1: there are no feasible points above and to the right). Similarly, in the ultimatum game an allocation of one cent to the Responder and $99.99 to the Proposer is also Pareto efficient. There is no way to make the Responder better off without making the Proposer worse off.

The same is true of problems such as the allocation of food between people who are more than satisfied and others who are starving. A very unequal distribution of food can be Pareto efficient as long as all the food is eaten by someone who enjoys it even a little.

In contrast, imagine how an engineer might evaluate a situation in which some people had barely enough food to survive while others got fat. The engineer might say: "This is not a sensible way to provide nutrition. It is clearly inefficient".

But the engineer would be using the everyday meaning of the term. *Pareto efficiency has nothing to do with whether the outcome is sensible*.

So while, in principle, Pareto inefficient allocations can and should be improved upon (by shifting to the allocation that is better for at least one and not worse for any). There may nevertheless be something wrong with many Pareto efficient allocations.

5.3 EVALUATING INSTITUTIONS AND OUTCOMES: FAIRNESS

For this reason we also evaluate allocations using the concept of *justice*: by which we mean, is it fair?

Suppose, in our ultimatum game, the Responder accepted an offer of one cent from a total of $100 (rather than refusing, and depriving the Proposer of $99.99). As we have seen in Unit 4 ultimatum game subjects in experiments around the world would typically reject such an offer, apparently because they judged it to be *unfair*. This would be the reaction of many of us if, instead of being subjects in the experimental

lab, we witnessed the two friends, An and Bai, walking down the street. Both spot a $100 bill, which An picks up and claims the right to distribute. An offers one cent to her friend Bai, and says she wants to keep the rest.

We might be outraged. But we might apply a different standard of justice if we found out that, though both An and Bai had worked hard all their lives, An had just lost her job and was homeless while Bai was well off. Letting An keep $99.99 might then seem fair. Thus we might apply a different standard of justice to the outcome of the game when we know all of the facts.

We could also apply a standard of fairness not to the outcome of the game, but to the *rules of the game* themselves. Suppose we had observed An proposing an even split, allocating $50 to Bai. Good for An, you say, that seems like a fair outcome. But if this had occurred because Bai had pulled a gun on An, and threatened that unless she offered an even split she would shoot her, we would probably judge the outcome to be unfair.

The example makes a basic point about fairness. Allocations can be judged unfair because of :

- *How unequal they are*: We measure this inequality along some dimension (such as income, or subjective wellbeing). These are *substantive judgments of fairness*.
- *How they came about*: Maybe by force, or by competition on a level playing field for example. These are *procedural judgements of fairness*.

Substantive and procedural judgements

The main difference between the two is that, to make a substantive judgement about fairness, all you have to know is the allocation itself. To make procedural evaluations we require knowledge of the rules of the game and other aspects of why this particular allocation occurred.

Two people making substantive evaluations of fairness about the same situation need not agree, of course. For example, they may disagree about whether fairness should be evaluated in terms of income or happiness. If we measure fairness using happiness as the criterion, a person with a serious physical or mental handicap may need much more income than a person without such disabilities to be equally satisfied with his or her life.

SUBSTANTIVE JUDGEMENTS

These are based on some measure of inequality in the allocation such as:

- *Income*: The reward in money, or some equivalent measure of the individual's command over valued goods and services
- *Happiness*: Measured by subjective wellbeing indicators, such as those introduced in Unit 1
- *Freedom*: The extent one can do (or be) what one chooses without socially imposed limits

DISCUSS 5.1: SUBSTANTIVE FAIRNESS

Consider the society you live in, or another society with which you are familiar.

1. How would you rate income, happiness and freedom as candidates for something that should be made more equal, to make the society fairer?
2. Are there other things that should be more equal to achieve greater fairness in this society?

PROCEDURAL JUDGEMENTS

These are based on an evaluation of the rules of the game that brought about the allocation including:

- *Voluntary* exchange of private property acquired by legitimate means: Were the actions resulting in the allocation the result of freely chosen actions by the individuals involved, for example each person buying or selling things that they had come to own through inheritance, purchase, or their own labour? Or was fraud or force involved?
- *Equal opportunity for economic advantage:* Did people have an equal opportunity to acquire a large share of the total to be divided up, or were they subjected to some kind of discrimination or other disadvantage because of their race, sexual preference, gender, or who their parents were?
- *Deservingness:* Did the rules of the game that determined how much each would get take account of the extent to which an individual worked hard, or otherwise upheld valued social norms?

DISCUSS 5.2: PROCEDURAL FAIRNESS

Consider the society in which you live, or another society with which you are familiar. How would your society score in the above procedural judgements of fairness?

Evaluating fairness in outcomes

We can use these differing judgements to evaluate an outcome in the ultimatum game. The experimental rules of the game will appear to most people's minds as procedurally fair:

- *Proposers were chosen randomly*
- The game was played anonymously: Discrimination could not have been involved
- All actions were voluntary: The Responder could refuse to accept the offer, and the Proposer is typically free to propose any amount

When Responders rejected Proposers' offers, they were objecting to the allocation itself: those who later said that a low offer was "unfair" were making a judgement about the outcome, and not about the rules of the game. This is a substantive, not a procedural judgement.

The rules of the game in the real economy are a long way from the fair procedures of the ultimatum game, and procedural judgements of unfairness are very important to many people, as we will see in Unit 19.

People's values about what is fair differ. Some, for example, regard any amount of inequality as fair, as long as the rules of the game that determine the allocation are procedurally fair. Others judge an allocation to be unfair if some people are seriously deprived of basic needs, while others consume luxuries.

The American philosopher John Rawls (1921-2002) devised a way to think about these disagreements that may clarify these arguments, and may even sometimes lead us to find common ground on questions of values. We follow three steps:

1. *Fairness applies to all people*: For example, were we to substitute the positions of An and Bai in the above example, so that it was Bai instead of An who picked up the $100 and made the proposal, it would not alter whether the outcome is fair or not.
2. *Imagine a veil of ignorance*: Because of this we can think about justice as if we were evaluating the rules of the game and the resulting outcomes from behind what Rawls called a veil of ignorance. By this he meant that we do not yet know which

positions we would occupy in the society we are considering: we could be male or female, healthy or ill, rich or poor (or with rich or poor parents), in a dominant or an "excluded" ethnic group, and so on. In the $100 on the street game, we would not know if we were the person picking up the money, or the person responding to the offer.

3. *How does the veil of ignorance affect your evaluation*? When we are behind this veil of ignorance, Rawls argued, we would evaluate the constitutions, laws, inheritance practices, and other institutions of a society as an impartial outsider.

The advantage of Rawls' veil of ignorance is that it invites you, in making a judgement about fairness, to put yourself in the shoes of a person who is quite different from who you really are. It does this by asking you to imagine that following your "choice" of a set of institutions, you would then become part of the society you have endorsed, but with an equal chance of having any of the positions occupied by individuals in that society.

DISCUSS 5.3: THE VEIL OF IGNORANCE

Suppose that behind a Rawlsian veil of ignorance you could choose to live in a society in which one (but only one) of the three procedural standards for fairness (voluntary exchange of property, equality of opportunity, and deservingness) would be the guiding principle for how the rules of the game are organised.

1. Which one would you choose?
2. Give reasons for your choice.

Neither philosophy, nor economics, nor any other science, can eliminate disagreements about questions of value. But economics can clarify:

- *How the dimensions of unfairness may be connected*: For example how the rules of the game that give special advantages to one or another group may affect the degree of inequality.
- *The trade-offs between the dimensions of fairness*: For example, do we have to compromise some of the equality of income conception of fairness in order to implement more of the equality of opportunity conception?
- *Public policies that may address concerns about unfairness*: It can also evaluate whether these policies compromise other objectives.

In the remainder of this unit we explore situations in which we describe who produces what, who gains from the process, and what they gain. Like the experiments in Unit 4, we will see that both cooperation and conflict occur. As in the experiments, and in history, we will find that the rules matter.

To do this, recall the model in Unit 3 of the farmer, Angela, who produces a crop. We will use a simple economic model in which two characters appear in a sequence of scenarios:

1. Initially, *Angela works the land on her own*, and as in Unit 3, gets everything she produces.
2. *Next, we introduce a second person who does not farm—but would also like some of the harvest*. He is called Bruno.
3. Initially, Bruno *can force Angela to work for him*. In order to survive, she has to do what he says.
4. Later, the rules change: *the rule of law replaces the rule of force*. Bruno can no longer coerce Angela to work. But he owns the land and if she wants to farm his land, she must agree, for example, to pay him some part of the harvest.
5. Eventually, the rules of the game have changed again in Angela's favour: Angela and her fellow farmers now have the right to vote and *legislation has been passed that increases Angela's claim on the harvest*.

For each of these steps we will use an economic model to analyse the changes, analysing them from the standpoint of both efficiency and the distribution of income between Angela and Bruno. Remember that:

- Economics can tell you whether an outcome is Pareto efficient or not.
- *But economics cannot tell you if this outcome is fair.* This depends on your own analysis of the problem using the concepts of substantive and procedural fairness.

5.4 MUTUAL GAINS AND CONFLICT

Recall that Angela's harvest depended on the amount of labour devoted to farming according to the production function. She worked the land, enjoyed the remainder of the day as free time, and consumed the grain that this activity produced. Recall also that the slope of the feasible consumption frontier is the *marginal rate of transformation* (MRT) of free time into grain.

Angela values both the grain produced and her free time, and the value that she places on each depends on how much of each she has. We represent these values, as we did in Unit 3, as indifference curves, giving all of the combinations of grain and

free time for which she does not prefer one to the other. Recall that the slope of the indifference curve is called the *marginal rate of substitution* (MRS) between grain and free time.

The steeper the indifference curve, the more Angela values free time compared to how much she values grain. You can see this in Figure 5.3 if you look at how steep the indifference curves are when she has 8 bushels of grain, and differing amounts of free time: as her free time increases (moving to the right) the curves become flatter. She values free time less.

We have drawn Angela's indifference curves in Figure 5.3 as we did in Figure 3.21 so that as she gets more grain the marginal rate of substitution does not change. You can see this in the figure by noticing that, at 16 hours of free time the curves are equally steep when she has less than 8 bushels, exactly 8 bushels and more than 8 bushels (the three indifference curves we have drawn).

What this means is that she values grain some constant amount relative to free time, independently of how much she has. (Why might this be? Probably because she does not eat it all, she sells it and uses the proceeds to buy other things she needs.) This is just a simplification that makes our model easier to understand. Remember: when drawing the indifference curves for the model in this unit, simply shift them up and down, keeping the MRS constant at a given amount of free time.

Angela is free to choose her typical hours of work to achieve her most preferred combination of free time and grain.

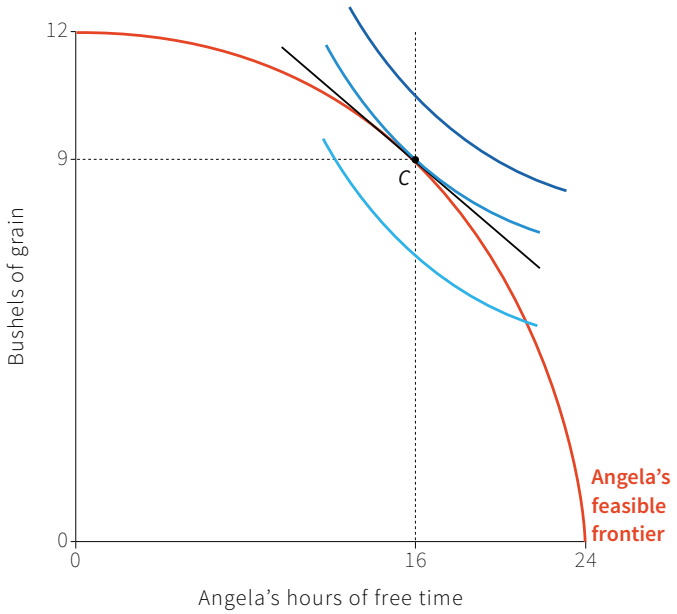


Figure 5.3. *Farmer-owner Angela's feasible frontier, best feasible indifference curve and choice of hours of work.*

Figure 5.3 shows that the best Angela can do, given the limits set by the feasible frontier, is to work for 8 hours, which gives her 16 hours of free time and production and consumption of 9 bushels of grain. This is the number of hours of work at which the marginal rate of substitution is equal to the marginal rate of transformation. She cannot do better than this! (If you're not sure why, go back to Unit 3 and check.) This Leibniz shows how to determine the best she can do using calculus, and our third Leibniz illustrates these *quasi-linear preferences* with specific utility and production functions.

But now, Angela has company. The other person is called Bruno, who is not a farmer but he will claim some of Angela's harvest. We will study a number of different rules of the game that explain how much is produced by Angela, and how it is divided between her and Bruno. For example, in one scenario, Bruno is the landowner and Angela pays some grain to him as rent for the use of the land.

Figure 5.4 shows Angela and Bruno's combined feasible frontier. The frontier indicates how many bushels of grain Angela can produce given how much free time she takes. For example, if Angela takes 12 hours free time and works for 12 hours, then she produces 10.5 bushels of grain. One possible outcome of the interaction between Angela and Bruno is that 5.25 bushels go to Bruno, and Angela retains the other 5.25 bushels for her own consumption.

The slideline demonstrates that each point in the figure is an allocation, showing how much work Angela did and how much grain she and Bruno got.

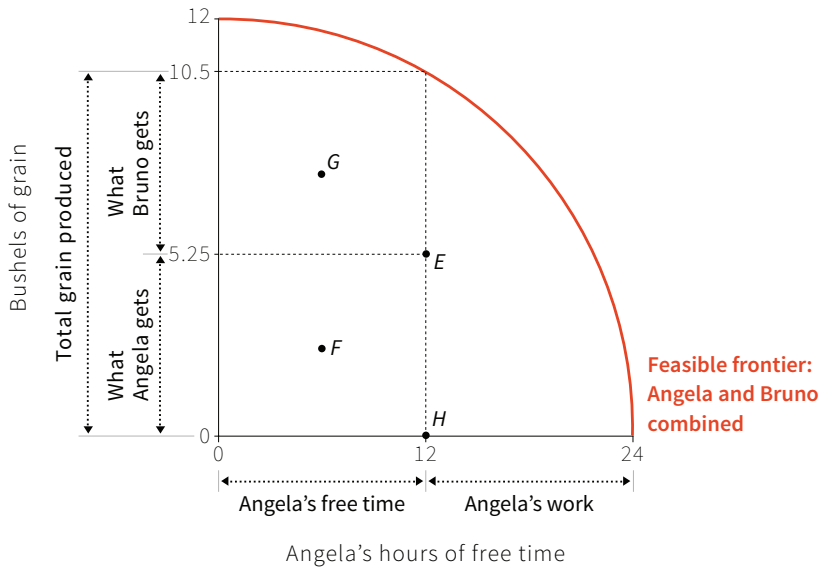


Figure 5.4 *Feasible outcomes in a bargaining problem.*

Which allocations are likely to occur? Not all of them are even possible. For example point *H* (the last step in Figure 5.4) is an allocation in which Angela works 12 hours a day and receives nothing (Bruno takes the entire harvest), so Angela would not survive. Of the allocations that are at least possible, the one that will occur depends on the rules of the game.

DISCUSS 5.4: USING INDIFFERENCE CURVES

In Figure 5.4, point *F* shows an allocation in which Angela works more and gets less, and point G shows the case in which she works more and gets more.

Show how you could use Angela's indifference curves (from Unit 3) to determine which of *E*, *F* or *G* she prefers.

Mutual gains and their distribution in Amazon Mechanical Turk

To see how every economic interaction, and the distribution of the gains, among people can be analysed, consider an example. When you sign up to Amazon Mechanical Turk, the online "marketplace for work," you may select one of more than 400,000 "human intelligence tasks", or HITs. The HITs are offered by Requesters. These are businesses and individuals looking for individuals to complete tasks such as listing product preferences, writing product descriptions, choosing preferred images, or even naming their honeymoon destination. Each HIT is described, along with the qualifications required and the payment per selected task. As a Mechanical Turk worker (a *turker*), you complete the HIT you have selected, and you are then paid.

The allocation in this case is the time you spent, your product transferred to the requester (the completion of the HIT), and your pay. The table below compares Angela's experience of farming with turking:

	FARMER	TURKER
EXTENT OF MUTUAL GAINS REALISED	Could both Angela and Bruno be better off if she worked fewer or more hours?	Is there a redesign of the HIT, the pay and other aspects so that both the requester and the turker could be better off?
DISTRIBUTION OF MUTUAL GAINS	Given her hours of work and his ownership rights in the land is the distribution of grain and her working time fair? Is his ownership of the land fair?	Is the pay a fair compensation for the effort and skills of the turker, given the benefit to the requester and their economic situation?

DISCUSS 5.5: AN ALLOCATION YOU HAVE KNOWN

Think of another job that you, or someone you know, has done (for example a barista or an office clerk). Using the analysis of the farmer and the turker above:

1. Identify the allocation associated with the job you have chosen.
2. Is the allocation Pareto efficient?
3. Is the allocation fair?

5.5 TECHNICALLY FEASIBLE ALLOCATIONS

Initially Angela could consume (or sell) everything she produced. Now Bruno has arrived, and he has a gun. He has the power to implement any allocation that he chooses. He is even more powerful than the dictator in the dictator game (in which a Proposer dictates how a given pie is to be divided). Why? Bruno can also determine the size of the pie, as well as how big his and Angela's slices are going to be.

Unlike the experimental subjects in Unit 4, in this model he is entirely self-interested. He wants only to maximise the amount of grain he can get from Angela. Angela, we will also assume, is similarly interested only in her own free time, and the grain she gets (as described by her indifference curves).

We now make another important assumption. If Angela does not work Bruno's land, he gets nothing (there are no other prospective farmers that he can exploit). What this means is that Bruno's reservation option (what he gets if Angela does not work for him) is zero. As a result, Bruno thinks about the future and he will not take so much grain that Angela will die. He has to impose some allocation that keeps her alive.

First, we will work out the set of technically feasible combinations of Angela's hours of work and the amount of grain she produces. By *technically feasible* we mean, the outcomes that are possible if the only limits on what can occur are the technology (the production function) and biology (Angela's need to get at least enough nutrition to carry out the work tasks of the allocation and survive).

How do we determine what is technically feasible? We already know from the feasible frontier in Figure 5.4 that the total amount consumed by Bruno and Angela combined cannot exceed the amount produced, which in turn depends on the hours that Angela works.

But we have also seen there are some combinations of grain and free time that would leave Angela so undernourished or overworked that she would not survive—for example point H in Figure 5.4 is *biologically infeasible*.

Angela's minimum nutrition requirements depend on how much she works. In Figure 5.5 we show the minimum amount of grain that would allow Angela to survive for each amount of work that she does. If she does not work at all, then she needs two bushels and a half to survive. If she expends energy working she needs more food; that's why the curve rises from right to left as her hours of work and higher expenditure of calories increase. This is the biological survival constraint. Points *below* it are biologically infeasible, while points above the feasible frontier are technically infeasible. The slope of the biological survival constraint is the marginal rate of substitution between free time and grain in securing Angela's biological survival.

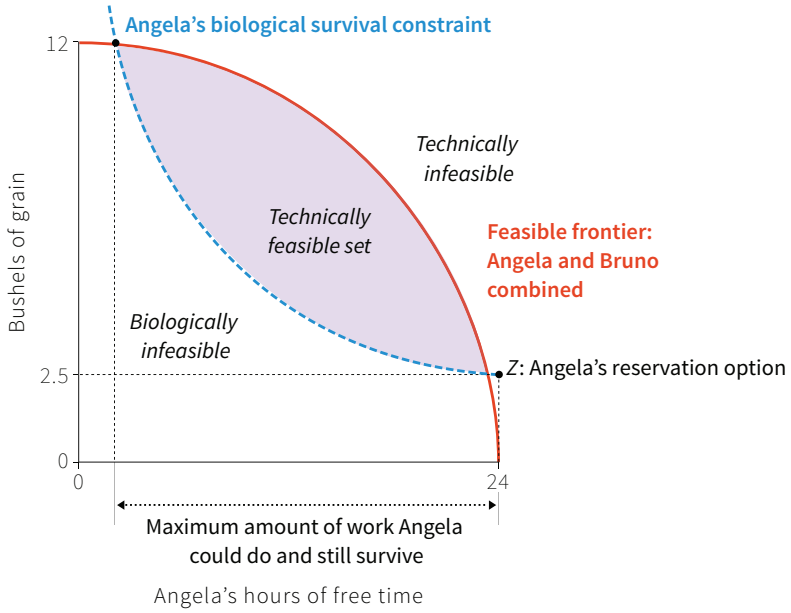


Figure 5.5 *Technically feasible allocations.*

Note that there is a maximum amount of work that would allow her barely to survive (because of the calories she burns up working). As we saw in Unit 2, throughout human history people crossed the survival threshold when the population outran the food supply. This is the logic of the Malthusian population trap. The productivity of labour placed a limit on how large the population could be.

DISCUSS 5.6: A GOOD HARVEST

Try shifting one or both of the curves in Figure 5.5 to represent the effects of a good harvest, or of population growth as in the Malthusian model in Unit 2. For example, how would you represent the Irish famine?Identify the allocation associated with the job you have chosen.

In Angela's case, however, it is not only the limited productivity of her labour that might jeopardise her survival, but also how much of what she produces is successfully claimed by Bruno. To see the difference note in Figure 5.5 that, if Angela could consume everything she produced, (the height of the feasible frontier) and if she could choose her hours of work, her survival would not be in jeopardy. The reason is that the biological survival constraint is below the feasible frontier for a great many hours of work and free time that she might choose. The question of biological feasibility arises because of Bruno's claims on her output.

In Figure 5.5, the boundaries of the feasible solutions to the allocation problem are formed by the feasible frontier and the biological survival constraint. This lens-shaped shaded area gives the technically possible outcomes. We can now ask what actually happens—which allocation occurs, and how does this depend on the institutions governing how Bruno and Angela interact?

5.6 ALLOCATIONS IMPOSED BY FORCE

With the help of his gun, Bruno can choose any point in the lens-shaped technically feasible set of allocations. But which will he choose?

He reasons like this:

Bruno For any number of hours that I order Angela to work, she will produce the amount determined by the feasible frontier of the production function. But for that amount of work I'll have to give her at least the amount shown by the biological survival constraint. I get to keep the difference between what she produces and what I need to give her, so that I can continue to exploit her. Therefore I should find the hours of Angela's work for which the vertical distance between the feasible frontier and the biological survival constraint (Figure 5.6) is the greatest.

Remember the amount that Bruno will get if he implements this strategy is his *economic rent*, in this case meaning the amount he gets over what he would get if Angela were not his slave (which, in this model, we set at zero).

Bruno first considers letting Angela continue working 8 hours a day, as she did when she had free access to the land. At that time she produced 9 bushels. If she works for 8 hours she needs 3.5 bushels of grain to survive. So Bruno could now take 5.5 bushels without jeopardising his future opportunities to benefit from Angela's labour.

Bruno is studying Figure 5.6 and asks for your help. You have worked out that the MRS is less than the MRT at 8 hours of work:

You Bruno, your plan cannot be right. If you forced her to work a little more, you wouldn't have to let her have much more grain because the biological survival constraint is relatively flat at 8 hours of work. But the feasible frontier is steep, so while you'd have to let her have a little more so that she would have the energy to work longer, she'd produce a lot more if you imposed longer hours..

You demonstrate the argument to him using the slideline in Figure 5.6, which indicates that the vertical distance between the feasible frontier and the biological survival constraint is the greatest when Angela works for 11 hours. If Bruno commands Angela to work for 11 hours then she will produce 10 bushels and Bruno will get to keep 6 bushels for himself. We can use Figure 5.6 to find out how many bushels of grain Bruno will get for any technically feasible allocation.

Look again at the last step in Figure 5.6. We have plotted the grey arrows showing the amount of grain that Bruno gets in the lower panel. You can see that the amount he gets falls when Angela works for more or less than 11 hours. By joining up the points we see that the amount Bruno gets is hump-shaped, and peaks at 11 hours of work and 13 hours of free time.

How can you calculate Angela's hours of work to give Bruno the greatest amount of grain, consistent with Angela surviving? Find out how to do this mathematically in this unit's Einstein section.

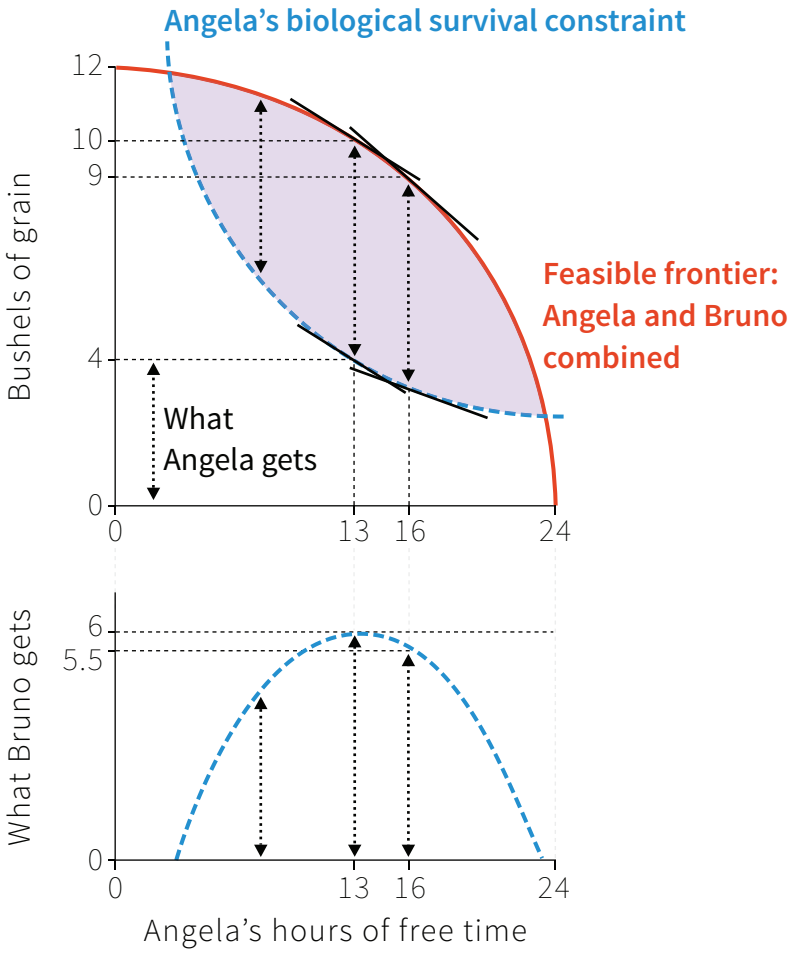


Figure 5.6 *Coercion: The maximum technically feasible transfer from Angela to Bruno.*

5.7 POWER AND THE DISTRIBUTION OF ECONOMIC RENTS

Once we move from studying a situation of coercion to one in which there is a legal system that prohibits Angela's enslavement and protects private property and the rights of landowners and workers, it is necessary to define new concepts.

In Unit 1, we defined private property as the right to use and exclude others from the use of something, and the right to sell it (or to transfer these rights to others). Bruno owns the land and

POWER

The ability to do and get the things we want in opposition to the intentions of others.

can exclude Angela if he chooses. But how much grain he will get as a result of his private ownership of the land still depends on his power over Angela, just as it did when he was able to use force.

Power in economics takes two main forms:

- *Setting the terms of an exchange* by making a take-it-or-leave-it offer (as in the ultimatum game).
- *Imposing or threatening to impose heavy costs,* unless the other acts in a way that serves the advantage of the person with power (as in Bruno's use of force).

The Proposer's advantage in the ultimatum game is a form of power called *bargaining power*. The Proposer makes a take-it-or-leave-it offer. The Responder could instead have been the lucky one, having the power associated with being Proposer, and going home with a half or more of the pie. In experiments the assignment of the role of Proposer or Responder, and hence the assignment of bargaining power, is usually done by chance.

In real economies, as in the case of Angela and Bruno, the assignment of power is definitely not random.

Those who own a factory or other business are the ones proposing the wage and the hours of work. Those seeking employment are like Responders, but with a difference: typically more than one person is seeking to land a job, so if one of them rejects the employer's proposal the employer can move to the next potential worker, just as in the ultimatum game when there are two Responders. There is another difference. Because the place of employment is the employer's private property, the employer can exclude the worker by firing her unless her work is up to the specifications of the employer.

We will see in the next unit how the labour market, along with other institutions, gives both kinds of power to employers. In Unit 7 we explain how firms selling goods sometimes have the power to propose take-it-or-leave-it offers to consumers, and in Unit 11 how the credit market gives power to banks and other lenders over people seeking mortgages and loans.

As we have seen in the previous section, power can be used in less subtle ways too. Perhaps the reason that Bruno is able to coerce Angela to work is that Angela's family was displaced from the land by force. In this case Bruno gained his power to make take-it-or-leave-it offers because he exercised a more overt form of power: physical coercion.

But, in a capitalist economy in a democratic society, most economic interactions are not conducted at the point of a gun. They are usually the result of individuals pursuing their objectives as best they can, given the property available to them, and given the power they exercise under the institutions of that economy. Recall from

Unit 4 that people bargain over their economic rent. Rents are also sometimes called *gains from exchange*, because they are how much a person gains by engaging in the exchange compared to not engaging.

When people participate voluntarily in an interaction, the sum of the rents is termed the *surplus* (or sometimes the *joint surplus*, to indicate that it's the sum of all of the rents).

Each person involved has to receive at least some rent, or otherwise they would have no incentive to participate. Angela, in the previous example, was coerced to farm Bruno's land. Next we look at the situation where she can simply say no. Angela is no longer a slave. Bruno has lost the power to coerce her, but not the power to make a take-it-or-leave-it offer, just like the Proposer in the ultimatum game.

5.8 BARGAINING POWER AND THE DISTRIBUTION OF THE SURPLUS

We check back on Angela and Bruno, and immediately notice that Bruno is now wearing a suit, and is no longer armed. He explains that this is no longer needed because there is now a government with laws administered by courts, and professional enforcers called the police. Bruno now owns the land, and Angela must have permission to use his property. He can offer a contract in which she can farm the land; in return she gives him part of the harvest. Alternatively, she can refuse the offer.

Bruno It used to be a matter of power, but now both Angela and I have property rights: I own the land, and she owns her own labour. The new rules of the game mean that I can no longer force Angela to work. She has to agree to the allocation that I propose.

You And if she doesn't?

Bruno Then it's no deal. She doesn't work on my land, I get nothing, and she gets barely enough to survive from the government.

Just like the ultimatum game, you think.

You So you and Angela have the same amount of power?

Bruno Certainly not! I am the one who gets to make a take-it-or-leave-it offer. I am like the Proposer in the ultimatum game; except that this is no game. If she refuses she goes hungry, and I have plenty of grain from other farmers.

You But if she refuses you get zero?

Bruno That never happens.

Why does he know this? Bruno knows that Angela, unlike the subjects in the ultimatum game experiments, is entirely self-interested (she does not punish an unfair offer). If he makes an offer that is just a tiny bit better for Angela than not working at all and getting subsistence rations, she will accept it.

Now he asks you a question similar to the one he had asked earlier:

Bruno In that case, what should my take-it-or-leave-it offer be?

You had answered by showing him the biological survival constraint. Now the limitation is not that the offer is such that Angela survives, but rather that she *agrees*. Over years of interacting with Angela and people like her he knows that she values her free time, so the more hours he offers her to work, the more he is going to have to pay.

You Why don't you just look at Angela's indifference curve that passes through the point where she does not work at all and barely survives? That will tell you how much is the least you can pay her for each of the hours of free time she would give up to work for you.

Point Z in Figure 5.7 is the allocation in which Angela does no work and gets only survival rations from the government (or from her family). This is her *reservation option*: if she refuses Bruno's offer she has this option in reserve. Use the slideline to see Angela's reservation curve: the curve giving all of the allocations that are just as highly valued by Angela as the reservation option. Below or to the left of the *reservation indifference curve* she is worse off than in her reservation option. Above and to the right she is better off.

The set of points bounded by the reservation indifference curve and the feasible frontier is the set of all economically feasible allocations, once Angela has to accept agree to the proposal that Bruno makes. Bruno thanks you for this handy new tool for figuring out the most he can get from Angela. You remind him that what you showed him is exactly the same as Angela's reservation indifference curve in Unit 3.

The biological survival constraint and the reservation indifference curve have a common point (Z): at that point, Angela does no work and gets subsistence rations from the government or her family. Other than that the two curves differ. The reservation indifference curve is uniformly above the biological survival constraint. The reason, you explain to Bruno, is that however hard she works along that frontier, she barely survives; and the more she works the less free time she has, so the unhappier she is. Along the reservation indifference curve, by contrast, she is just as well off as at her reservation option, meaning that being able to keep more of the grain that she produces compensates exactly for her lost free time.

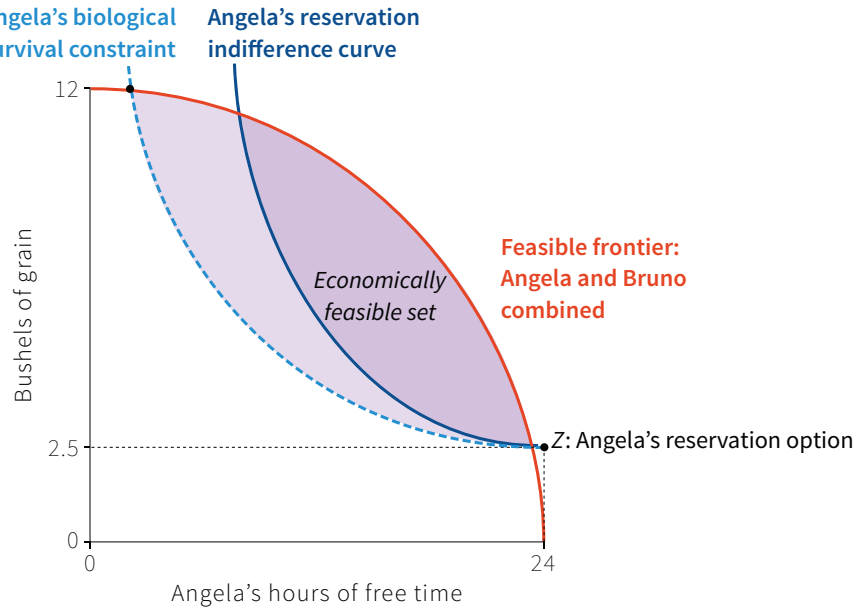


Figure 5.7 *Economically feasible allocations when exchange is voluntary.*

DISCUSS 5.7: BIOLOGICAL FEASIBILITY

Using Figures 5.5 and 5.7:

1. Explain why a point on the biological survival constraint is higher (more grain is required) when Angela has fewer hours of free time. Why does the curve also get steeper when she works more?
2. Explain why the biologically feasible set does not have to be equal to the economically feasible set.
3. Explain (by shifting the curves) what happens if a more nutritious kind of grain is available to Angela.
4. Explain (by shifting the curves) what happens if there is a famine due to a bad harvest.
5. What happens in Figure 5.5 when there is population growth as in the Malthusian model from Unit 2?

We can see that both Angela and Bruno may benefit if a deal can be made. The reason is that their exchange—allowing her to use his land (that is, not using his property right to exclude her) in return for her sharing some of what she produces—makes it possible for both to be better off than if no deal had been struck. Both benefit from a deal:

- As long as Bruno gets some of the crop he will do better than if there is no deal.
- As long as Angela's share makes her better off than she would have been if she took her reservation option, taking account of her work hours, she will be also benefit.

This potential for mutual gain is why their exchange need not take place at the point of a gun, but can be motivated by the desire of both to be better off. All of the economically feasible allocations that represent mutual gains are in the last step of Figure 5.7.

Of course the fact that mutual gains are possible does not mean that both Angela and Bruno will benefit equally. It all depends on the institutions in force. If Bruno has the power to make a take-it-or-leave-it offer, subject only to Angela's agreement, he can capture the entire surplus (less the tiny bit necessary to get Angela to agree). Bruno knows this already.

Once you have explained the reservation indifference curve to him, Bruno knows which allocation he wants. He maximises the amount of grain he can get consistent with Angela farming the land at the maximum height of the lens-shaped region. This is the maximum vertical distance between Angela's reservation indifference curve and the feasible frontier, which will be where the MRT on the feasible frontier is equal to the MRS in Angela's reservation indifference curve. Figure 5.8a shows that this allocation requires Angela to work for fewer than she did under coercion.

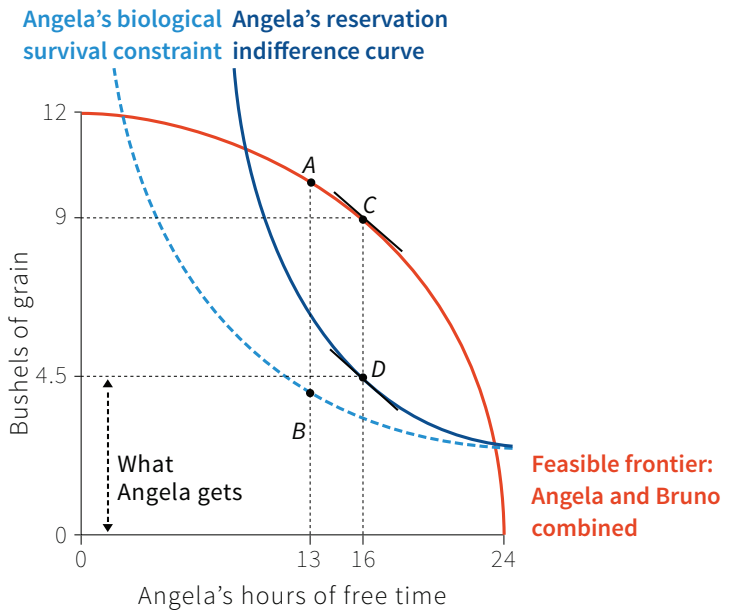


Figure 5.8a *Bruno's take-it-or-leave-it proposal when Angela can refuse.*

So Bruno would like Angela to work for 8 hours and give him 4.5 bushels of grain. How can he implement this allocation? All he has to do is to make a take-it-or-leave-it offer of a contract allowing Angela to work the land in return for a land rent of 4.5 bushels per day. If Angela has to pay 4.5 bushels (CD in Figure 5.8a) then she

will choose to produce at point C, where she works for 8 hours. You can see this in the figure; if she produced at any other point on the feasible frontier and then gave Bruno 4.5 bushels she would have lower utility—she would be below her reservation indifference curve. But she can achieve her reservation utility by working for 8 hours, so she will accept the contract.

Since Angela is on her reservation indifference curve, only Bruno benefits from this exchange. The entire surplus goes to Bruno. His economic rent (equal to the land rent she pays him) is the surplus.

Notice that Angela chooses the same hours of work under this contract that she did when she could work the land without paying rent. Why does this happen? However much rent Angela has to pay, she will choose her hours of work to maximise her utility, so she will produce at a point on the feasible frontier where the MRT is equal to her MRS. And we know that her preferences are such that her MRS doesn't change with the amount of grain she consumes, so it will not be affected by the rent. This means that if she can choose her hours, she will work for 8 hours irrespective of the land rent (as long as this gives her at least her reservation utility).

Figure 5.8b shows how the surplus varies with Angela's hours if she is on her reservation indifference curve, in comparison with the case of coercion, when she was on her biological survival constraint. Bruno should make an offer where Angela has to pay him 4.5 bushels to rent the land. As we can see, she will work for 8 hours and keep the difference between the total amount she produces and the rent that she paid to Bruno. We can see that the amount of grain Bruno gets falls as Angela works more or less than 8 hours. Again, if we join up the points we can see that the amount Bruno gets is hump-shaped. The peak is lower than when Bruno has the power to order Angela to work, because he needs Angela to agree to the proposal.

DISCUSS 5.8: TAKE IT...

1. Why is it Bruno, and not Angela, who is in a position to make a take-it-or-leave-it offer?
2. Are there conditions under which the farmer, not the landowner, might have this power?

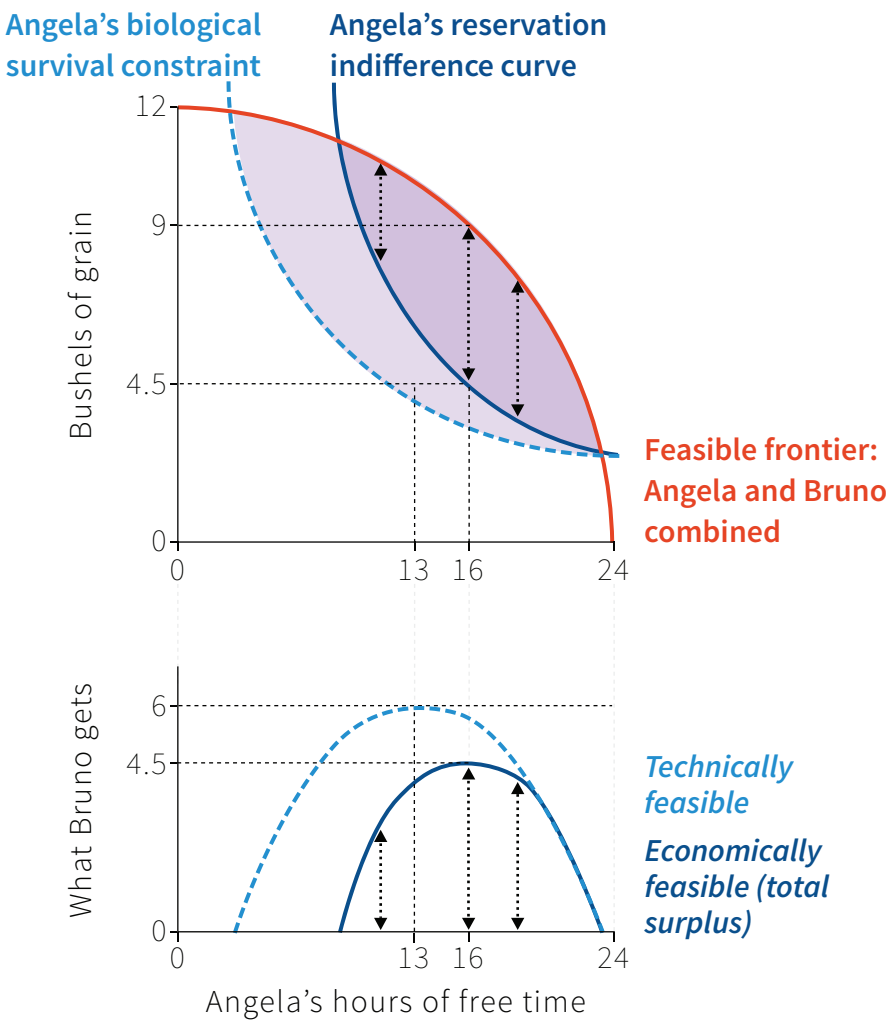


Figure 5.8b *Bruno's take-it-or-leave-it proposal when Angela can refuse.*

DISCUSS 5.9: ... OR LEAVE IT

We have assumed that Angela would accept anything that made her better off than her reservation position, irrespective of how unfair she thought the division of the surplus was. But suppose that Angela was like the players in the ultimatum game experiment who preferred to give up their rent entirely, rather than participate in an unfair division of the surplus with the Proposer.

What would her new reservation indifference curve look like?

5.9 THE PARETO EFFICIENCY CURVE FOR DIFFERING DISTRIBUTIONS OF THE SURPLUS

Remember that Angela chose to work for 8 hours, producing 9 bushels of grain, when she had to pay rent and when she did not. In both cases there is a surplus of 4.5 bushels: the difference between the amount of grain produced, and the amount that would give Angela her reservation utility.

The two cases differ in who gets the surplus: when Angela had to pay rent to Bruno, he took the whole of the surplus; when she could work the land for herself she received the surplus herself. But both allocations are *Pareto efficient*:

- They are Pareto efficient because, at both, the MRT on the feasible frontier is equal to the Angela's MRS.
- Any allocation where the MRT is not equal to the MRS would not be Pareto efficient, because Angela's hours could be then changed to make her better off without affecting what Bruno gets.

Figure 5.9a shows that there are many other Pareto-efficient allocations for which Angela works for 8 hours but the surplus is distributed differently. Point C is the outcome when Angela is an independent farmer. Use the slideline to compare this with Bruno's take-or-leave-it offer, and to see the other Pareto-efficient allocations.

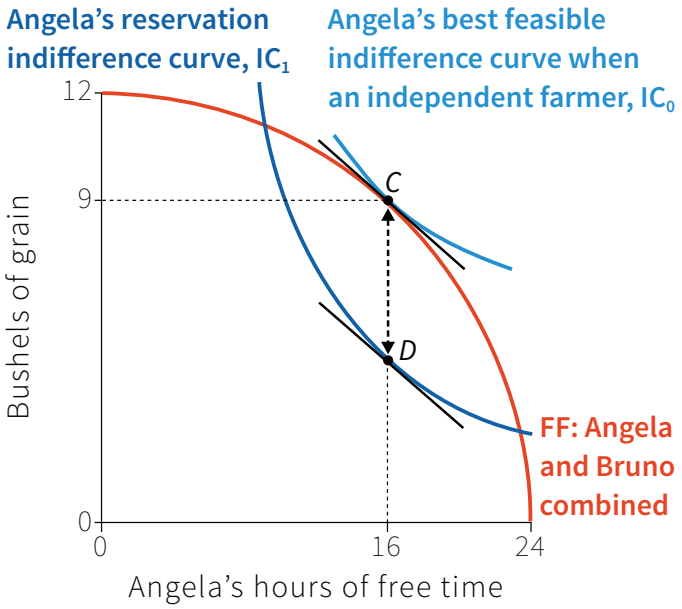


Figure 5.9a *Pareto efficient allocations and conflicts of interest.*

What is true of *C* and *D* is true of each point between them: the distribution of the surplus differs but the allocations are Pareto efficient. All of the points that are Pareto efficient make up what is called the *Pareto efficiency curve*. (You will also hear it called the *contract curve*, even in situations where there is no contact, which is why we prefer the more descriptive term Pareto efficiency curve.) Points *C*, *D*, and the points in between are thus on the Pareto efficiency curve.

PARETO EFFICIENCY, AND THE PARETO EFFICIENCY CURVE

We know that a *Pareto efficient* allocation is:

- An allocation with the property that there is no alternative technically feasible allocation in which at least one person would be better off, and nobody worse off.
- The set of all such allocations is the *Pareto efficiency curve*. It is also referred to as the contract curve, even in social interactions in which there is no contract.

In Figure 5.9b, we look at a hypothetical allocation where Angela shares in the surplus by getting grain over and above her reservation indifference curve.

In such an allocation, Angela's economic rent is *GD* and Bruno's is *CG*: the sum is the surplus, *CD*. Because *MRT* = *MRS* at this point (like at points *C* and *D*), the allocation is Pareto efficient.

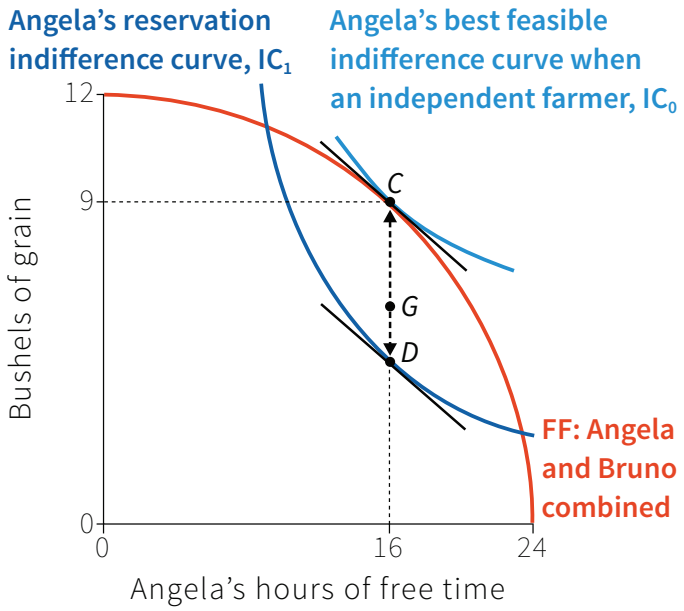


Figure 5.9b *Angela's economic rent, the land rent she pays to Bruno, and Bruno's economic rent along the Pareto efficiency curve.*

Both Bruno and Angela prefer Point *G* to their reservation option, point *Z* in Figure 5.7, in which Bruno gets zero and Angela gets 2.5 bushels and 24 hours of free time. This is true of every point in the dark shaded lens-shaped area shown in the last step of Figure 5.7. Each point in this space is called a Pareto improvement over point *Z* (the mutual reservation option) and is said to *Pareto dominate* point *Z*.

The Pareto efficiency curve between points *C* and *D* is important because it gives all of the points that are both a Pareto improvement over the "no deal" reservation option, and are Pareto efficient. If Bruno and Angela are going to bargain over who gets what, they obviously should be thinking about settling on a point on the line *CD*. If they end up anywhere else, as we will soon see, they both could do better if they kept on bargaining.

5.10 POLITICS: SHARING THE SURPLUS

Bruno thinks that the new rules, requiring him to make an offer that Angela will not refuse, are not so bad after all. Angela too is better off than she had been when she had barely enough to survive. But she would like a share in the surplus.

She and her fellow farm workers agitate for a new law that limits the work time that can be imposed to 4 hours a day, while requiring that total pay is at least 4.5 bushels. They threaten to not work at all unless the law is passed.

Bruno Angela, you and your colleagues are bluffing.

Angela No we are not: we would be no worse off at our reservation option than under your contract, working the hours and receiving the small fraction of the harvest that you impose!

Angela and her fellow workers win, and the new law limits the working day to 4 hours.

How did things work out?

Angela had been working for 8 hours and getting 4.5 bushels of grain, before the short hours law (when Bruno was charging maximum rent). This is point *D* in Figure 5.10. The new law implements the allocation in which Angela and her friends now work 4 hours, getting 20 hours of free time as a result and the same number of bushels. Since they have the same amount of grain and 4 more hours of free time, they are better off. Figure 5.10 shows they are now on a higher indifference curve..

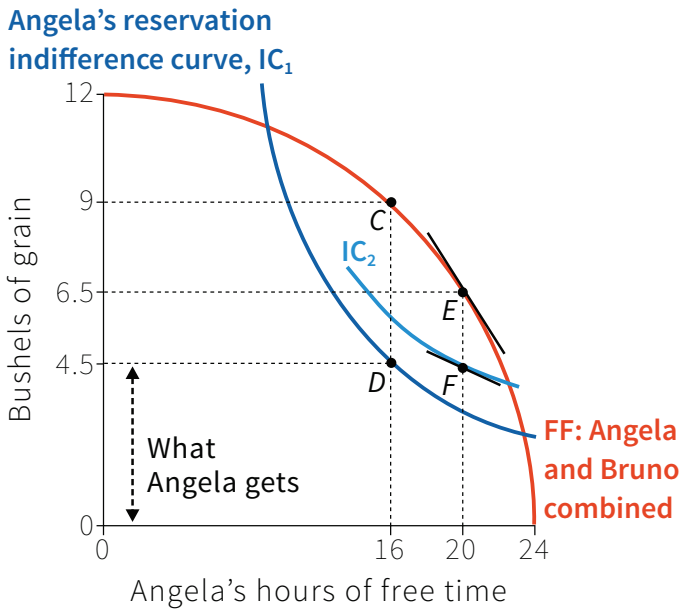


Figure 5.10 *The effect of an increase in Angela's bargaining power through legislation.*

You can see that Angela is better off at *F* than at *D*. She is also better off than she would be with her reservation option, which means she is now receiving an economic rent. The size of Angela's rent is a measure in bushels of grain units of how much better off she is at the new situation (*F*) than at her reservation option (on IC_1). It can be described in a number of ways:

- *The vertical distance between her reservation indifference curve and the indifference curve she is able to achieve under the new legislation*, measured in grain. In Figure 5.10, this is the vertical distance between indifference curves IC_1 and IC_2.
- *The maximum amount of grain per year that Angela would give up to live under the new law* rather than in the situation before the law was passed.
- (Because Angela is obviously political) *the amount she would be willing to pay so that the law passed*, for example by lobbying the legislature or contributing to election campaigns.

Bruno is not happy. You try to cheer him up:

You Next time they threaten to quit working, saying they have nothing to lose, they really *will* be bluffing. They have their economic rent to lose. Remember, Bruno, Angela's economic rent is for her the opportunity cost of telling you to get lost: now she will not be tempted to walk away.

Bruno can see that this rent will be useful later. We will return to it in the next unit, and later in the course.

5.11 BARGAINING TO A PARETO-EFFICIENT SHARING OF THE SURPLUS

Angela and her friends are pleased with their success. She asks what you think of the new policy.

You Congratulations, but your policy is far from the best you could do.

Angela Why?

You Because you are not on the Pareto efficiency curve! Under your new law, Bruno is getting the amount of bushels shown by *EF* and cannot make you work more than four hours. So why don't you offer to continue to pay him the same amount of bushels that he is now getting, in exchange for agreeing to let you keep anything you produce above the amount you have given to him? Then you get to choose how many hours you work.

The small print in the law allows a longer work day if both parties agree, as long as the 4-hour day is the workers' reservation option if no agreement is reached.

You now redraw Figure 5.10 and use the concepts of the surplus and the Pareto efficiency curve from Figure 5.9 to show Angela how she can get an even better deal.

You Look at Figure 5.11. The surplus is largest at 8 hours of work, just like in Figure 5.8b. When you work for 4 hours the surplus is small, and you pay most of it to Bruno. If you increase the surplus you can pay him the same amount, and your own surplus will be bigger—so you will be better off. Use the slideline to see how.

The move away from point *D* (at which Bruno had all the bargaining power and experienced all the gains from exchange) to a point where Angela is better off consists of two distinct steps:

1. From *D* to *F*, the outcome imposed by Angela's legislation. This was definitely not win-win: Bruno lost because his economic rent at *F* is less than the maximum feasible rent that he got at *D*. Angela benefitted.
2. Once at the legislated outcome, there were many win-win possibilities open to them. They are shown by the segment *GH*, on Pareto efficiency curve. Win-win alternatives to the allocation at *F* are possible by definition, because *F* was not Pareto efficient.

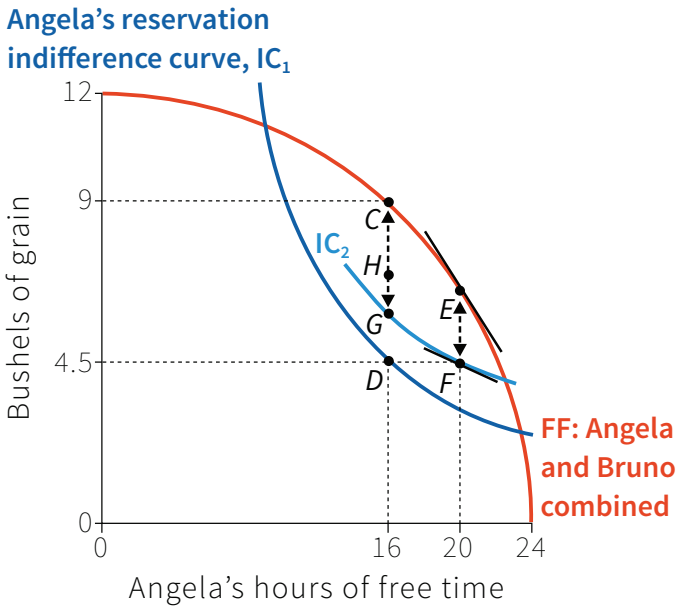


Figure 5.11 *Bargaining to restore Pareto efficiency.*

Bruno wants to negotiate. He is not happy with Angela's proposal of *H*.

Bruno I am no better off under this new plan than I would be if I just accepted the legislation that the farmers passed.

You But Bruno, Angela now has bargaining power, too. The legislation changed her reservation option, so it is no longer 24 hours of free time at survival rations. Her reservation option is now the legislated allocation at point *F*. I suggest you make her a counter offer.

Bruno Angela: I'll let you work the land for as many hours as you choose if you pay me half a bushel more than *EF*.

They shake hands on the deal.

Because Angela is free to choose her work hours, subject only to paying Bruno the extra half bushel, she will work 8 hours where *MRT* = *MRS*. Because this deal lies between *G* and *H*, it is a Pareto improvement over point F. Moreover because it is on the Pareto efficient curve *CD*, we know there are no further Pareto improvements to be made starting from any point on the line segment *GH*. This is true of every other allocation on *GH*—they differ only in the distribution of the mutual gains, as some favour Angela while others favour Bruno. Where they end up will depend on their bargaining power.

5.12 A POLICY TO REDISTRIBUTE THE SURPLUS AND RAISE EFFICIENCY

Angela and Bruno live in the hypothetical world of an economic model. But real farmers and landowners face similar problems.

In the Indian state of West Bengal, landless farmers rent land from landowners, and give them a share of the crop as payment. A farmer working under this kind of contract is called a sharecropper, or a *bargadar* in the Bengali language.

The contractual arrangements throughout this vast state—home to more people than live in Germany—varied little from village to village, with virtually all bargadars giving half their crop to the landowner at harvest time. This had been the norm since at least the 1340s, when Ib'n Battuta had visited Bengal on his travels.

But, like Angela, in the second half of the 20th century many thought this was unfair, because of the extreme levels of deprivation among the sharecroppers in West Bengal: in 1973, 73% of the rural population lived in poverty, one of the highest poverty rates in India. In 1978 the newly-elected Left Front government of West Bengal adopted new laws, called *Operation Barga*. This stated that:

- Bargadars could keep up to three-quarters of their crop.
- Bargadars were protected from eviction by landowners, provided they met this 25% quota.

Both provisions of Operation Barga were advocated as a way of increasing efficiency. There are certainly reasons to predict that the size of the pie would increase, as well as the incomes of the farmers:

- *Bargarders had a greater incentive to work hard and well*: Keeping a larger share meant that that there was a greater reward if they grew more crops.
- *Bargarders had an incentive to invest in improving the land*: They had confidence that they would farm the same plot of land in the future, so would be rewarded for their investment.

West Bengal enjoyed a subsequent dramatic increase in farm output per unit of land, as well as farm incomes. By comparing the output of farms before and after the implementation of Operation Barga, economists concluded that both effects—improved work motivation and investment—occurred: one study suggested that the Operation Barga was responsible for around 28% of the subsequent growth in agricultural productivity in the region. The empowerment of the bargadars also had positive spillover effects as local governments became more responsive to the needs of poor farmers.

Operation Barga was later cited by the World Bank as an example of good policy for economic development.

But the limitation on the crop share that bargaders needed to give up lowered the incomes of some landowners. Therefore the change in policy was *not* a Pareto improvement.

Even though it was not Pareto efficient, by increasing the income of the poorest people in West Bengal, we might judge that Operation Barga was fair. We can assume that many people in West Bengal thought so, because they continued to vote for the Left Front alliance. It stayed in power from 1977 until 2011.

We can study the effect of land reform on the distribution of income using the Lorenz curve introduced in Unit 1 to compare inequalities among people earning varying incomes. We do not have detailed information for Operation Barga, but we can consider inequalities in a hypothetical village with just two groups of people: sharecroppers and landowners.

Imagine there are 10 landowners, each owning 10 hectares, and 90 others who farm the land as sharecroppers, but who own no land. The Lorenz curve for the distribution of land is given by the lower boundary of the shaded area in Figure 5.12, indicating that the poorest 90% of the population own none of the land, while the remaining 10% own all of the land.

If instead each member of the population owned one hectare of land—perfect equality in land ownership—then the Lorenz curve would be a line at a 45-degree angle (indicating that the "poorest" 10% of the population have 10% of the land, and so on, but we say "poorest" in inverted commas because if everyone had one hectare, then everyone is equally poor, and equally rich).

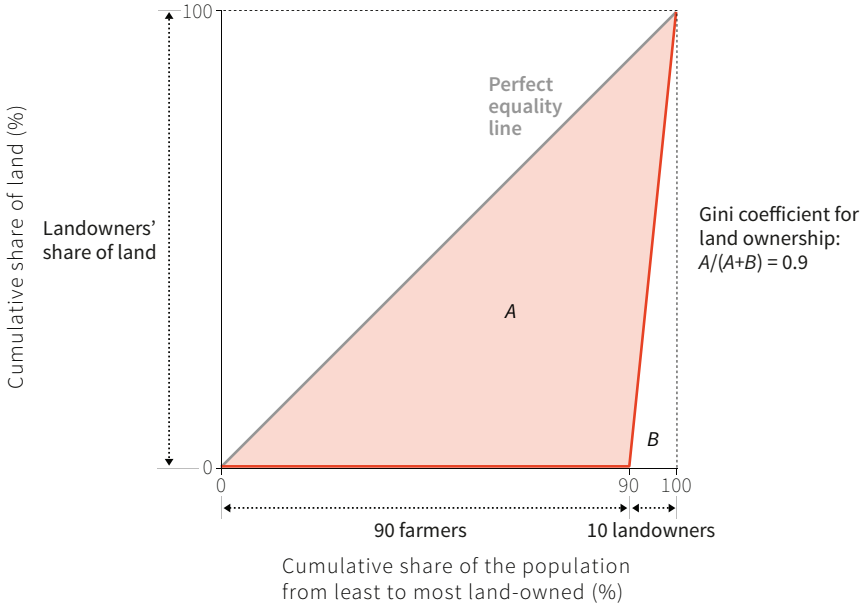


Figure 5.12 *The Lorenz curve and Gini coefficient for wealth ownership.*

In the economy depicted in Figure 5.12, the Gini coefficient for land ownership is 0.9.

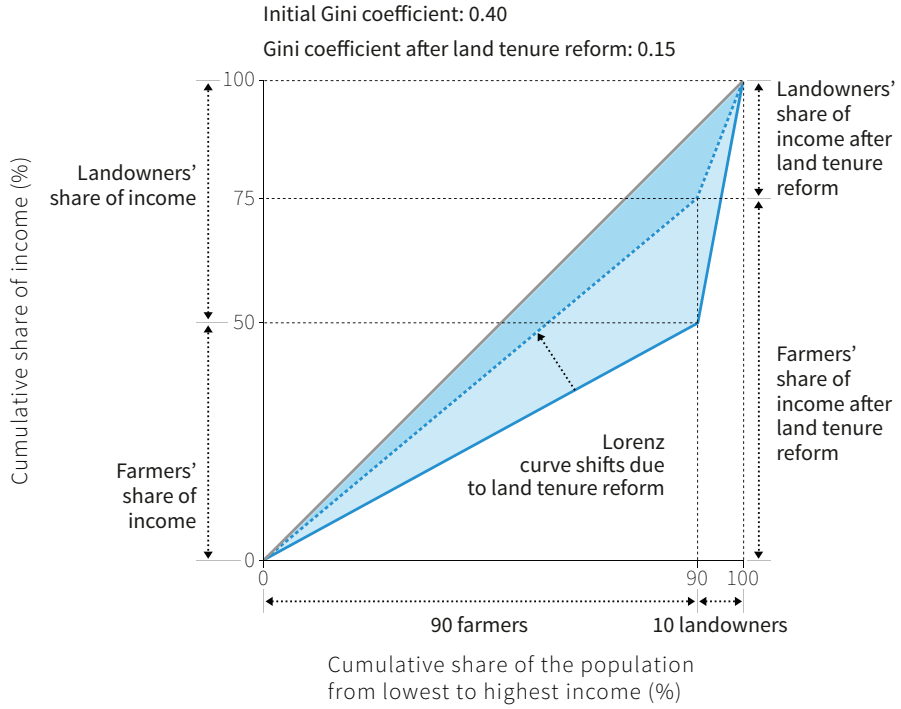


Figure 5.13 *How a land tenure reform in West Bengal, India reduced the Gini coefficient.*

To see how Operation Barga worked, recall that farmers are paying a rent of 50% of their crop to the landowner. Operation Barga raised the farmer's crop share as shown in the figure, so that the farmers now received 75% of the crop. As a result, the Gini coefficient of income was reduced from 0.4 (similar to the US) to 0.15 (well below that of the most equal of the rich economies, such as Denmark).

5.13 CONCLUSION

When the pirates on Captain Roberts' *The Rover* agreed unanimously to a constitution, they accepted a set of rules of the game—that is, institutions—that would determine who did what on the ship and how the spoils were to be divided. The same is true for requesters and turkers who sign up for Amazon Mechanical Turk.

When two or more people voluntarily come together to undertake a common project, whether pirates, turkers, or Angela farming Bruno's land (when Bruno's proposed terms were at least minimally acceptable to Angela), their cooperation results in

the possibility of mutual gains from exchange. They are potentially both better off having engaged in a common project than they would have been otherwise, because otherwise they gain only their reservation utility.

The same is true when people directly exchange, or buy and sell, goods for money. If you have more apples than you can consume, and your neighbour has an abundance of pears, the same logic applies. The apples are worth less to you than they are to your neighbour, and the pears are worth more to you. So there must be some rate of exchange under which you are happy to exchange some apples for some pears.

This logic applies to land and labour, or requesters with tasks and would-be turkers with time on their hands. When people with differing needs, property and capacities meet, there is an opportunity to generate gains for all of them. That is why people often like to come together in markets, online exchanges or pirate ships. The mutual gains are the pie.

Whether they are able to mutually benefit depends on technology and biology. If Bruno's land had been so unproductive that no amount of labour would have produced enough to compensate Angela for her time, then there would have been no deal that they could strike. Amazon Mechanical Turk is successful because people around the world (many of them with abundant free time and little income) can work on the projects of comparatively rich, but busy, requesters in the US.

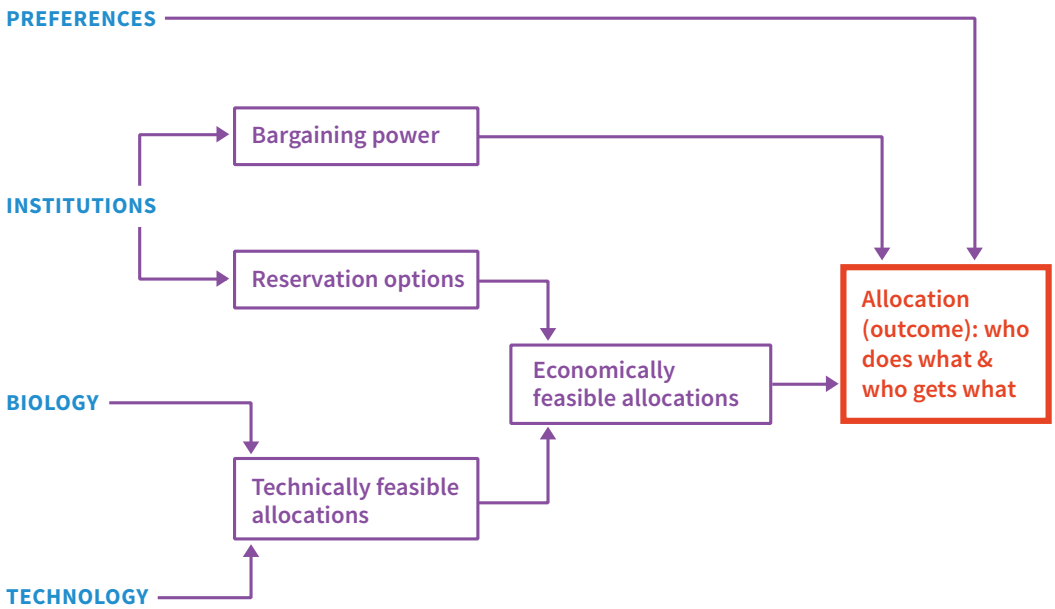


Figure 5.14 *Institutions, mutually beneficial interactions and distribution.*

Among the set of allocations that are technically feasible, the ones that we observe through history are largely the result of the institutions, including property rights and bargaining power, that were (or are) present in the economy. The institutions answer two questions (summarised in Figure 5.14):

- *Who does what*, so that mutual gains are possible?
- *Who gets what*, or how are the mutual gains distributed among the parties to the exchange?

Figure 5.15 summarises the two ways to evaluate the allocations introduced in sections 5.3 and 5.4. Of course, the Pareto efficiency and fairness of the allocation are not the only values we might use to evaluate economic interactions. If we value the freedom of the participants we might also be concerned about the process: could they refuse to participate without fear of physical harm or other substantial costs? We might also value interactions that help people learn and adhere to other values that society holds to be important, such as tolerance, honesty and generosity.

Our imagined story of Angela and Bruno, and the true story of the Bengali bargadars, teach three lessons, to which we will return in subsequent units when we discuss policies to try to implement Pareto efficient outcomes with distributions that are considered to be fair by most people.

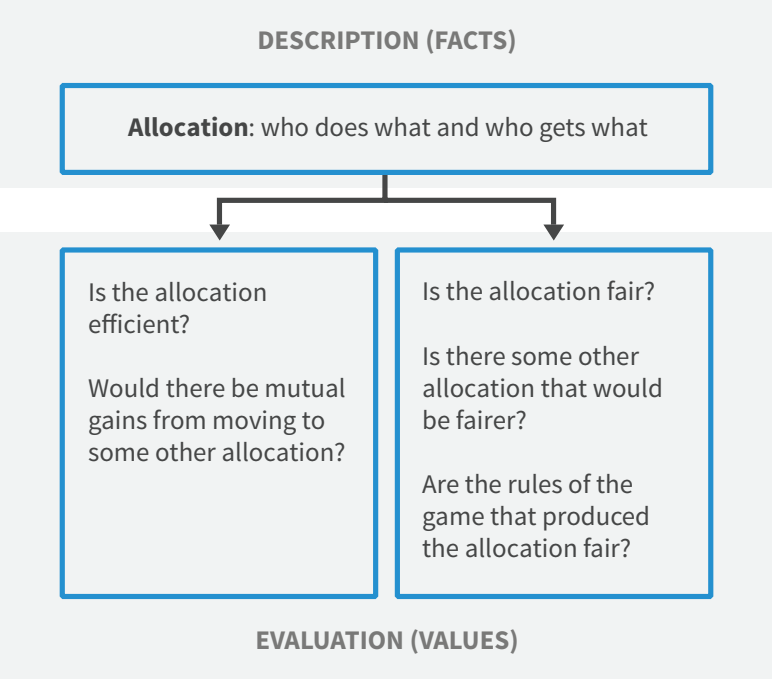


Figure 5.15 *Efficiency and fairness.*

- *When one person or group has power to dictate the allocation, subject only to not making the other party worse off than in their reservation option, the powerful party will capture the entire surplus.* They will implement an allocation that makes their gains as large as possible, subject only to the other party not being worse off than with no exchange at all. If they have done this, then there cannot be any way to make either of them better off without making the other worse off. So the result must be Pareto efficient!

- *Those who consider their treatment unfair often have some power to influence the outcome* through legislation and other political means, and the result may be a more fair distribution in their eyes or ours, but one that is not necessarily Pareto efficient. So societies may face trade-offs between Pareto efficient but unfair outcomes, and fair but Pareto inefficient outcomes.
- *If we have institutions under which people can jointly deliberate, agree on, and enforce alternative allocations then there may be outcomes that make both parties better off, and that are also fairer than the status quo.* Angela and Bruno managed this. Starting from a very uneven distribution of the benefits of exchange (only Bruno benefitted), legislation was passed increasing Angela's bargaining power, and then the two privately agreed on a win-win outcome that was Pareto efficient. When this is possible, we need not accept the trade-off between Pareto efficiency and fairness. A combination of legislation and private bargaining resulted in a Pareto efficient allocation, along with redistribution of rents towards the least well off.

These lessons make it clear that institutions are important for determining both the efficiency and the fairness of economic allocations. Aside from the government and families, the most important institutions of a modern economy are markets and firms. In the next unit, we study how firms (business organisations) address questions of allocation. We need to know how they work, and how well they work.

In the units that follow we study markets. Some have single firms selling to many buyers; others have large numbers of buyers and sellers. We investigate how they may implement allocations that allow mutual gains from exchange, and how they influence how those gains are distributed among buyers, sellers, and others. We also study the combination of markets and firms that make up the modern capitalist economy, and ask how successfully this set of institutions allows for Pareto efficient allocations and fair distributions.

CONCEPTS INTRODUCED IN UNIT 5

Before you move on, review these definitions:

- Surplus
- Power
- Bargaining power
- Pareto efficiency
- Pareto efficiency curve
- Economic rent (compared to land rent)
- Substantive and procedural concepts of fairness

In later units we will ask what we can do to make the results closer to being Pareto efficient—and also more fair.

Key points in Unit 5

Technically feasible

Technology and biology limit the allocations that are technically feasible.

Economically feasible

Actors' reservation options, their preferences and economic institutions determine the subset of technically feasible allocations that are economically feasible.

Power

Two forms of power are the ability to set the terms of an exchange and to get others to act in one's interest by imposing (or threatening to impose) penalties if they do not.

Institutions affect power

Economic and political institutions (the rules of the game) affect actors' power and the allocations that result from social interactions.

Pareto efficiency

Pareto efficiency provides a valuable but incomplete evaluation of economic outcomes.

Substantive and procedural justice

The fairness of both allocations and the rules of the game that produced them may be evaluated using substantive and procedural concepts of justice.

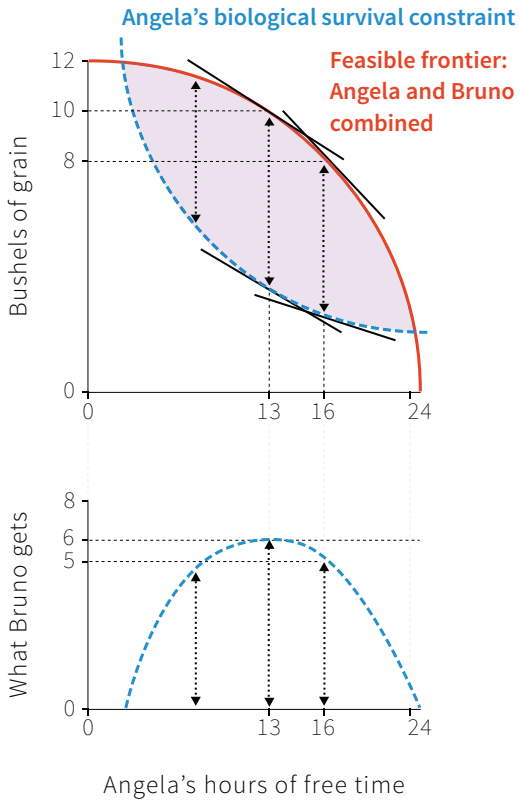
Efficiency and fairness

Both public policies and private bargaining among economic actors can contribute to outcomes that are Pareto efficient and also fairer.

5.14 EINSTEIN

Angela's hours of work

How can you figure out Angela's hours of work that gave Bruno the greatest amount of grain, consistent with Angela surviving? To the right of 13 hours (more free time for Angela), the biological survival constraint is flatter than the feasible frontier. That means the marginal rate of transformation of hours of labour into output is greater than the marginal rate of substitution of hours of labour into subsistence nutrition requirements. So moving to the left (Angela working more) results in an increase in production (her marginal product) that is greater than the increase in her subsistence needs. So Angela working more increases what Bruno gets, which you recall is his economic rent. To the left of 13 hours of free time (Angela working more), the reverse is true. Bruno's surplus is greatest at the hours of work where the slopes of the two frontiers are equal. That is:



marginal rate of transformation of work hours into grain output = *marginal rate of substitution work hours into subsistence requirements*

$$MRT = MRS$$

5.15 READ MORE

Bibliography

3. Alchian, Armen. 1987. 'Property Rights.' In *The New Palgrave Dictionary of Economics*, edited by John Eatwell, Murray Milgate, and Peter Newman. London: Macmillan.
4. Banerjee, Abhijit V., Paul J. Gertler, and Maitreesh Ghatak. 2002. 'Empowerment and Efficiency: Tenancy Reform in West Bengal.' *Journal of Political Economy* 110 (2): 239–80.

5. Bowles, Samuel, and Herbert Gintis. 2008. 'Power.' In *The New Palgrave Dictionary of Economics*, edited by Steven N. Durlauf and Lawrence E. Blume, 2nd ed. Basingstoke: Palgrave Macmillan.
6. Greif, Avner. 2006. 'History Lessons: The Birth of Impersonal Exchange: The Community Responsibility System and Impartial Justice.' *Journal of Economic Perspectives* 20 (2): 221–36.
7. Greif, Avner. 2007. *Institutions and the Path to the Modern Economy*. Cambridge: Cambridge University Press.
8. Juravich, Tom, and Kate Bronfenbrenner. 1999. *Ravenswood: The Steelworker's Victory and the Revival of American Labor* (ILR Press Books). Ithaca, NY: Cornell University Press.
9. Krueger, Alan B., and Alexandre Mas. 2004. 'Strikes, Scabs, and Tread Separations: Labor Strife and the Production of Defective Bridgestone/Firestone Tires.' *Journal of Political Economy* 112 (2): 253–89.
10. Leeson, Peter T. 2007. 'An–arrgh–chy: The Law and Economics of Pirate Organization.' *Journal of Political Economy* 115 (6): 1049–94.
11. Leeson, Peter T. 2009. *The Invisible Hook: The Hidden Economics of Pirates*. Princeton, NJ: Princeton University Press.
12. Murray, Dian H. 1987. *Pirates of the South China Coast, 1790-1810*. Stanford, CA: Stanford University Press.
13. Pareto, Vilfredo. (1906) 2014. *Manual of Political Economy: A Variorum Translation and Critical Edition*. Edited by Aldo Montesano, Alberto Zanni, and Luigino Bruni. Oxford, New York, NY: Oxford University Press.
14. Raychaudhuri, Ajitava. 2004. *Lessons from the Land Reform Movement in West Bengal, India*. Washington, DC: World Bank.
15. *The Economist*. 2013. 'More Sophisticated than You Thought.' November 2.

6

THE FIRM: OWNERS, MANAGERS AND EMPLOYEES

cc by B.D.'s world, Flickr

HOW THE INTERACTIONS AMONG THE FIRM'S OWNERS, MANAGERS AND EMPLOYEES INFLUENCE WAGES, WORK, AND PROFITS, AND HOW THIS AFFECTS THE WORKINGS OF THE ENTIRE ECONOMY

- The firm is an actor in the capitalist economy, and a stage on which interactions among the firm's employees, managers and owners are played out
- The balance of bargaining power among employees, managers and owners affects how the mutual gains created in the firm are distributed
- Hiring labour is different from buying other goods and services, and the contract between the employer and the employee does not cover what the employer really cares about: how hard and well the employee works
- Firms do not pay the lowest wages possible. Instead they set wages to motivate employees to work effectively, to stay with the firm, and to make it practical for the firms to recruit new workers when they need them
- The wages that firms pay their employees are influenced by the supply and demand for labour, and other factors that change the balance of bargaining power among the firm's actors
- The wage curve shows the relationship between wages and unemployment in the economy as a whole

Institute for New Economic Thinking
UCL SciencesPo Azim Premji University
FRIENDS PROVIDENT Foundation

Apple's iPhone and iPad are iconic American hi-tech products, yet neither is assembled in the US. Until 2011 a single company, Foxconn, produced every iPhone and iPad in factories in China, mainly so that Apple could take advantage of lower wages and other costs.

Apple Store, Fifth Avenue, New York.
Photo *by Jorge Lascar.*

The components of the iPhone and iPad for the most part do not come from China, but are sourced from around the world. Components such as the flash memory, display module and touch screen are made by companies including Toshiba and Sharp in Japan; the microprocessor by Samsung in South Korea; other components by Infineon in Germany. Like other firms, Apple makes profits by finding the supplier that can provide inputs at the least cost, whether the input is a component or labour, and wherever in the world that supplier may be located.

The cost of assembling the components into the final product in China is small—making up 4% of total cost—compared to the cost of components sourced from high-wage economies such as Germany and Japan. Almost half of Apple's employees in the US sell Apple products rather than making them, while firms compete on a global scale to win the lucrative business of supplying Apple with its components. The cost of producing the iPhone is far lower than the price Apple charges: in 2009, when the iPhone 3G cost $178 to manufacture, it was retailing in the US for $499.

Apple is not alone in outsourcing (or offshoring) production to countries that are not the main market for the goods produced. In most manufacturing industries, firms based in rich countries have transferred a significant proportion of production, previously done by local employees, to poorer countries where wages are lower. But Apple and other firms are looking for more than cheap labour: wages in some of Apple's source countries such as Japan are higher than in the US. Sometimes (in Germany, for example) much higher.

Other industries, particularly garment manufacturing, have relocated primarily to low wage economies. More than 97% of apparel and 98% of footwear sold in the US by American brands and retailers is made overseas. Garment manufacturing in the US is so rare that the company American Apparel can make this feature of its clothes a distinctive selling point. China, Bangladesh, Cambodia, Indonesia and Vietnam have become the world's main exporters of textiles and clothing. At the time of the Industrial Revolution, the world's largest exporter was Britain.

Also, in developing countries, additional business costs such as public holidays or health and safety rules are far lower, and environmental regulations are often less strict.

FIRM

Firms employ people and purchase the inputs they need to produce and market goods at prices that more than cover the cost of production.

Apple, Samsung, American Apparel and Toshiba are business organisations called *firms*. Not everyone is employed in a firm. For example, many farmers, carpenters, software developers or personal trainers work independently, as neither employee nor employer. Others work for governments and not-for-profit organisations; but the majority of people in rich nations make their living by working in a firm.

Firms are major actors in the economy and in this and the next unit we explain how firms work. Firms are often referred to as if they were a person: we talk about "the price Apple charges", and "American Apparel's choice of an advertising strategy". But while firms are actors, firms are also the stage on which the people who make up the firm—employees, managers, and owners—act out their sometimes common, sometimes competing, interests. In our video Richard Freeman of Harvard University and the National Bureau of Economic Research explains some of the consequences of outsourcing for these actors.

To understand the firm we will investigate a model of how the firm sets wages and how employees respond to these wages. In earlier units we introduced models that also help explain the importance of firms in how we live:

- In Unit 1 we defined economics as the study of how people interact with each other and with nature in producing our livelihoods, so *work is an important part of economics*.
- In Unit 2 we studied how in *an entire economy, population growth and wages were mutually determined* in the long run, and how this Malthusian link was broken with the capitalist revolution.
- In Unit 3 we concentrated on how *individuals choose hours of work*, and trade off free time into goods.
- In Unit 4, because *making a living is not something we determine alone*, we modelled how people who are motivated by some combination of self-interest and social preferences interact strategically—for example how much work to devote to maintaining the community's irrigation system.
- In Unit 5, to *clarify how conflicts and mutual gains arise in economic interactions*, we introduced a model of bargaining between two individuals determining work hours and the division of a surplus.

Each of these models illuminates some aspects of the economy, while setting aside others. In Unit 2 we did not study how the length of the working day was determined while the population and economy were growing. In Unit 3 we did not model

how the wage or the marginal rate of transformation of free time into goods was determined when we modelled a decision on working hours. In Unit 2 we told a story of conflicting interests over wages, but we did not model strategic interaction and bargaining until Units 4 and 5. And in Unit 5 we used the story of just two (imaginary) people called Bruno and Angela to model how bargaining may affect the Pareto efficiency and fairness of allocations.

Here we begin our study of the modern capitalist economy, modelling how this bargaining takes place in the firm. We will see that:

- People who make up a firm realise mutual gains from exchange because they are all better off in their firm than they would be on their own, or in some other firm.
- They will also have conflicting interests about how these gains will be shared, just like Bruno and Angela.

In Unit 7, we look at the firm as an actor in its relationship with other firms and with its customers.

6.1 FIRMS, MARKETS AND THE DIVISION OF LABOUR

The economy is made up of people doing different things; some producing the Apple display modules, others producing American Apparel clothing. Among those producing the display modules there are also a vast number of distinct tasks, and different employees within Toshiba or Sharp, the companies that produce the modules for Apple, do these tasks.

When people engage in different tasks as part of the production of a given product, this is called the division of labour. There has to be a way to coordinate the division of labour because the results of their efforts have to get from the hands of the producers to those who use them. This was not a difficult problem when most families produced most of what they needed, relying little on other producers. But in a modern global economy, different products like shirts or flash drives, and different components of products, like the collars on the shirts or the microprocessors in the computers, have to end up where they are needed.

Setting aside the work done in families, in a capitalist economy the division of labour is coordinated in two major ways: firms and markets. Through firms, the components of the goods produced by different people in different departments of the firm, or even in different firms, are brought together to assemble the finished shirt or

iPhone. By buying and selling goods on markets, the finished iPhone gets from the producer into the pocket of the consumer, and the American Apparel shirt ends up on somebody's back.

So in this unit we study firms. In units to follow, we study markets. Herbert Simon, an economist, used the view from Mars to explain why it is important to study both:

GREAT ECONOMISTS

HERBERT SIMON

Imagine a visitor from Mars, Herbert "Herb" Simon (1916-2001) urged his readers. Equipped with a telescope that revealed social structure, what would our visitor see? Companies might appear as green fields, he suggested, divisions and departments as faint contours within. Connecting these fields, red lines of buying and selling. Within these fields, blue lines of authority, connecting boss and employee, foreman and assembly-worker, mentor and mentee.

Traditionally, economists had focused on the market and the competitive setting of prices. But to a visitor from Mars, Simon suggested:

"Organisations would be the dominant feature of the landscape. A message sent back home, describing the scene, would speak of 'large green areas interconnected by red lines.' It would not likely speak of 'a network of red lines connecting green spots'."
— Herbert Simon, *Organizations and Markets* (1991)

Trained as a political scientist, Simon's desire to understand society led him to a study of both institutions and the human mind—to open the "black box" of motivations that economists had come to take for granted. He was celebrated in departments of computer science, psychology, and, of course, economics, for which he won the Nobel prize in 1978.

A firm, he pointed out, is not simply an agent, shifting to match supply and demand. It is composed of individuals, whose needs and desires might conflict. In what ways could these differences be resolved? Simon asked when an individual would shift from contract work (a "sale" of a particular, predefined task) to an employment relation (where a boss dictates the task after the sale—the relationship at the heart of a firm)?

When the desired task is easy to specify in a contract, Simon explained that we could view this as simply work-for-hire. But high uncertainty (the employer not knowing in advance what needed to be done) would make it impossible to specify contractually what the worker was to do and, in this case, the result would be an employer-employee relation that is characteristic of the firm.

This early work showcased two of Simon's lasting interests: the complexity of economic relations, where one might sell an obligation that was incompletely described, and the role of uncertainty in changing the nature of decision-making. His argument demonstrated the emergence of the "boss".

Understanding how contract work turns into employment implies only that we understand a particular relationship between two members of an organisation. We have yet to explain the firm as a whole—the Martian's green fields.

What makes a good organisation? This is a question for psychologists as much as econometricians, because we know that incentives that tie individual rewards to the success of the organisation appear to have little effect.

Simon's intellectual career can be contrasted with another great economist, Friedrich Hayek, whose ideas we will examine in detail in Unit 9. Both were interested in how societies could thrive in the face of uncertainty and imperfect agents. For Hayek, the price mechanism was all: a device to collect and process vast quantities of information, and so synchronise systems of arbitrary size.

But for Simon, the price mechanism needed to be supplemented—even supplanted—by institutions and governments better equipped to handle uncertainty and rapid change. These alternative "authority mechanisms" draw on partially-understood aspects of the human psyche: loyalty, group identification, and creative satisfaction.

By the time of his death in 2001, Simon had seen many of his ideas reach the mainstream. Behavioural economics has roots in his attempts to build economic theories that reflect empirical data. Simon's view from Mars shows that economics could not be self-contained science: an economist needs to be both a mathematician, working with decision-sets and utilities, and a social psychologist, reasoning about the motivations of human relationships.

Firms, along with private property and markets, are among the institutions that define a capitalist economy. The firm differs in an important respect from markets:

- *Markets* involve a decentralisation of power.
- *Firms* represent a concentration of economic power in the hands of the owners and managers.

The prices that motivate and constrain people's actions in a market are the result of the actions of thousands or millions of individuals, not a decision by someone in authority. The idea of private property specifically limits the things a government or anyone else can do with your possessions.

In a firm, by contrast, owners or their managers direct the activities of their employees, who may number in the thousands or even millions. The managers of Walmart, the world's largest retailer, order the activities of 2.2 million employees, a larger number of people than any army in world history before the 19th century. Walmart is an exceptionally large firm, but it is not exceptional in that it brings together a large number of people who work together in a way coordinated (by the management) to make profits.

Firms do not form spontaneously and then disappear like flashmobs. Like any organisation, firms have a decision-making process and ways of imposing their decisions on the people in it. Figure 6.1 is a simplified picture of the firm's actors and decision-making structure.

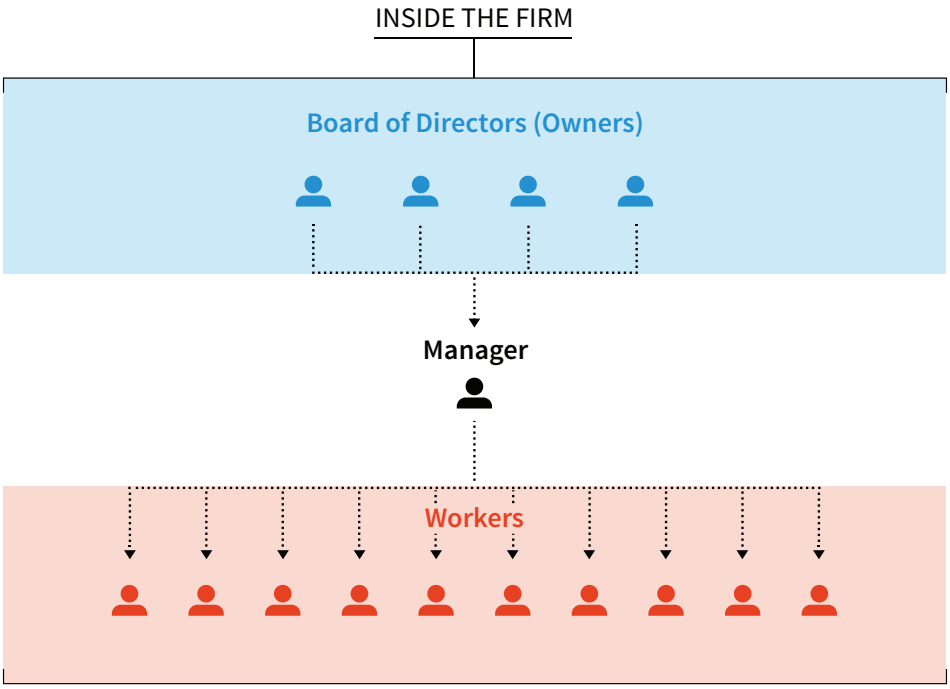


Figure 6.1 *The firm's actors and decision-making structures.*

The arrows in the figure represent directions or commands (which may or may not be carried out as intended). The owners, through their board of directors, direct the manager (or managers) to implement decisions about the long-term strategies of the firm concerning how, what, and where to produce.

The manager, in turn, assigns workers to the tasks required for these decisions to be implemented, and attempts to ensure that the assignments will be carried out. When we say that "Apple outsourced its component production" or "The firm sets a price of $10.75", we mean that the decision-making process in the firm resulted in these actions.

This relationship between the firm and its employees contrasts with the firm's relationship to its customers, which we study in the next unit. The bakery firm cannot text its customers to tell them to "Show up at 8am and purchase two loaves of bread at the price of €1 each". It could tempt its customers with a special offer but, unlike the employer with its employees, it cannot require them to show up. When you buy or sell something, it is generally voluntary. In buying or selling you respond to prices, not orders.

The firm is different: it is defined by having a decision-making structure in which some people have power over others. Ronald Coase, the economist who founded the study of the firm as both a stage and an actor, wrote:

"If a workman moves from department Y to department X, he does not go because of a change in prices but because he is ordered to do so… the distinguishing mark of the firm is the suppression of the price mechanism."
— Ronald H. Coase, *The Nature of the Firm* (1937)

Coase pointed out that the firm in a capitalist economy is a miniature, privately owned, centrally planned economy. Its top-down decision-making structure resembles the centralised direction of production in entire economies that took place in many Communist countries (and in the US and the UK during the second world war).

This difference between markets and firms is clear when we consider the differing kinds of *contracts* that form the basis of exchange.

A sale contract for a car transfers ownership, meaning that the new owner can now use the car and exclude others from its use. A rental contract on an apartment does not transfer ownership of the apartment (which would include the right to sell it); instead it allows the tenant to engage in a limited set of uses of the apartment, including right to exclude others (including the landlord) from its use.

CONTRACT

A *contract* is a legal document or understanding that specifies a set of actions that the parties to the contract are to undertake.

Under a *wage labour* contract, an employee gives the employer the right to direct him or her to be at work at specific times, and during those times to accept the authority of the employer over the use of his or her time.

The employer does not own the employee as a result of this contract. If the employer did, the employee would be called a slave. We might say that the employer has rented the employee for part of the day. To summarise:

- *Contracts for goods* sold in markets transfer ownership of the good from the seller to the buyer.
- *Contracts for labour* grant authority to direct the activities of the firm's employee from the employee to the manager or owner.

Firms differ from markets in another way: social interactions in firms sometimes extend over decades, or even a lifetime. In markets our contacts are typically shortlived and not repeated. In markets we shop around. One of the reasons for the difference is that working in a firm—as either a manager or an employee—means acquiring a network of associates who are essential to being able to carry out the job well. Some of our workmates will become our friends. Managers and employees also acquire both technical and social skills that are specific to the firm they work for.

Oliver Williamson, an economist, termed these skills, networks, and friendships *firm-specific assets* because, when employees leave the firm, they lose their value. Think about how different this is to the social interactions in the market: you often know the face or even the name of a person from whom you buy, or to whom you sell something; but the relationship is typically temporary, and so this knowledge often has no value.

This social fact becomes important economically when economic changes disrupt social interactions.

Imagine how your life as a shopper changes if your local grocery store closes tomorrow. You would have to find a new place to shop, and it might take you a few minutes to learn where the various items are on display.

Now imagine what would change if the company in which you work goes out of business tomorrow. You would lose your network of work associates, the friendships at work, and your firm-specific social and technical skills would overnight have become useless to you. You might have to move to a new town: your children would need to change school, and so they would lose all their friends too.

Thus there are two important differences between firms and markets:

- *Firm-specific assets*: Unlike shopping at a particular store, working in a firm means accumulating firm-specific assets that will be lost if the connection to the firm is severed.
- *Power*: Working in a firm, unlike buying or selling products in a market, means engaging in a relationship in which some individuals who have the power to issue orders to others, with the expectation that those orders will be carried out.

Owners and managers exercise power over employees. (In the next unit we will see that firms also exercise a different kind of power, called market power, when they set prices.)

The people making up the firm—owners, managers, and employees—are united in their common interest in the success of the firm because all of them would suffer if it were to fail. Their interests will clash about the distribution of wages, managerial salaries and owners' profits in a successful firm, as well as other policies such as conditions of work, managerial perks, and who makes key decisions—such as whether Apple should assemble iPhones in China or the US.

Why do some economic interactions take place in firms and some in markets? In firms, the manager's authority dictates the allocation; in markets, allocations result from a two-sided deal that one can easily walk away from. Coase showed that the "suppression of the price mechanism" within firms had cost advantages to the firm. So the firm constantly has to decide between two options: "make it" (produce a component itself) and "buy it" (from another firm, using markets). The boundaries of this divide between the firm and the market are set, Coase explained, by the relative costs of the "make it" and "buy it" options.

DISCUSS 6.1: THE STRUCTURE OF A FIRM

In Figure 6.1 we showed the actors and decision-making structure of a typical firm.

1. How do the actors and decision-making structure of Google, Wikipedia, and a family farm compare with this?
2. Redraw Figure 6.1 to represent these entities.

6.2 OTHER PEOPLE'S MONEY: THE SEPARATION OF OWNERSHIP AND CONTROL

The firm's profits legally belong to the people who own the assets of the firm, such as the capital goods. The owners direct the other members of the firm to take actions contributing to the firm's profits. This in turn will increase the value of the firm's assets, and improve the wealth of the owners.

They own whatever remains after revenues (the proceeds from sale of the products) are used to pay employees and managers, suppliers, creditors and taxes. Profit is the *residual*: that is, what's left of the revenues after these payments. The owners claim it, which is why they are called *residual claimants*. Managers (unless they are also owners) are not residual claimants. Neither are employees.

This has an important implication. A job done well by a manager or an employee, one that causes the firm's revenues to increase, will benefit the owners; but unless it results in a promotion, a bonus or a salary increase, *it will not benefit the actor*. This is one reason we consider the firm as a stage, and one on which not all the actors have the same interests.

In small enterprises, the owners are typically also the managers, in charge of operational and strategic decisions. As an example, consider a restaurant owned by a sole proprietor, who decides on the menu, hours of operation, marketing strategies, choice of suppliers, and the size and compensation of the workforce. In most cases the owner will try to maximise the profits of the enterprise by providing the kinds of food and ambiance the people want at competitive prices. Unlike Apple, the owner cannot outsource dishwashing or table service to a low-wage location.

In large corporations, there are typically many owners. Most of them play no part in management. The owners of the firm are the individuals and institutions, such as pension funds, that own the shares issued by the firm. By issuing shares to the general public a company can raise capital to finance its growth, leaving strategic and operational decisions to a relatively small group of specialised managers. These decisions include what, where and how to manufacture, or how much to pay employees and managers. The senior management of a firm is also responsible for deciding how much of the firm's profit is distributed to shareholders in the form of dividends, and how much is retained to finance growth. Of course the owners benefit from the firm's growth because what they own is part of the value of the firm, which increases as the firm grows.

When managers decide on the use of other people's funds, this is referred to as the *separation of ownership and control*.

The separation of ownership and control results in a potential conflict of interest. The decisions of managers affect profits, and profits decide the incomes of the owners. But it is not always in the interest of managers to seek to maximise profits. They may choose to take actions that provide them with benefits for themselves, at the expense of the owners. They may spend as much as possible on their company credit card, or seek to increase their own power, even if that is not in the interests of shareholders.

Even single owners of firms are not required to maximise their profits. Restaurant owners can choose menus they personally like, or waiters who are their friends. But, unlike managers, when they lose profits as a result, the cost comes directly out of their pocket.

In the 18th century Adam Smith observed the tendency of senior managers to serve their own interests, rather than those of shareholders. He had this to say of the managers of what were then called joint-stock companies:

"[B]eing the managers rather of other people's money than of their own, it cannot well be expected, that they should watch over it with the same anxious vigilance with which the partners in a [firm managed by its owners] frequently watch over their own… Negligence and profusion, therefore, must always prevail, more or less, in the management of the affairs of such a company."
— Adam Smith, *The Wealth of Nations* (1776)

Smith had not seen the modern firm; but he understood the problems raised by the separation of ownership and control. There are two ways that owners can incentivise managers to serve their interests. They structure contracts so that managerial compensation depends on the performance of the company's share price. Also the firm's board of directors, which represents the firm's shareholders and has a membership typically with a substantial share in the firm (like a representative of a pension fund), monitors the performance of the management. The board has the authority to dismiss managers, and shareholders in turn have the right to replace members of the board. The owners of large companies with many shareholders rarely exercise this authority, partly because shareholders are a large and diverse group that cannot easily get together to decide something. Occasionally, however, this free-rider problem is overcome and a shareholder with a large stake in a company may lead a shareholder revolt to change or influence senior management.

When we assume that firms always maximise profits we are making a simplification, but a reasonable one for most purposes:

- Owners have a strong interest in profit maximisation because it is the basis of their wealth.
- Market competition among firms penalises and even eliminates firms that do not make substantial profits for their owners. We saw this process in Unit 1 and Unit 2 as part of the explanation of the permanent technological revolution, and it applies to all aspects of the firms' decisions.

6.3 OTHER PEOPLE'S LABOUR

The firm does not solely manage, as Smith put it, "other people's money". The decision-makers in a firm decide on the use of other people's labour too: the efforts of their employees. People participate in firms because they can do better if they are part of the firm than if they are not. As in all voluntary economic interactions there

are mutual gains. But just as conflicts arise between owners and managers, there will generally be differences between owners and managers on the one hand, and employees on the other, about how the firm will use the strength, creativity and other skills of its employees.

A firm's profits (before the payment of taxes) depend on three things:

- Costs of acquiring the inputs necessary for the production process.
- Output: how much these inputs produce.
- Sales revenues received from selling goods or services.

In Unit 2 we saw how a firm might increase output without raising costs by adopting a new technology. In Unit 7 we study how the firm decides what price to charge. Here we study how firms seek to minimise the cost of acquiring the necessary labour to produce the goods and services that they sell.

Hiring employees is different to buying other goods and services. When we buy a shirt, or pay someone to mow a lawn, it is clear what we get for our cash. If we don't get it, we don't pay. If we have already paid, we go to court and get our money back. But a firm cannot write an enforceable employment contract specifying the exact tasks employees have to perform or not get paid.

This is true for three reasons:

- When the firm writes a contract for the employment of a worker, *it cannot know exactly what it will need the employee to do*, because this will be determined by necessarily unforeseen future events.
- It would be impractical or *too costly for the firm to observe* exactly how much effort each employee puts in to the job.
- Even if the firm somehow acquired this information, it *could not be the basis of an enforceable contract*.

To understand the last point, consider a restaurant owner, who would like his staff to serve customers in a pleasant manner. Imagine how difficult it would be for a court to decide whether the owner can withhold wages from a waitress because she had not smiled often enough.

In addition, because it costs the firm to find and train new workers, the employer has an interest in employees staying in the job once they have been selected and trained. But, in most countries,

INCOMPLETE CONTRACT

We say that the employment contract is *incomplete* because:

- It cannot protect the firm from employees who fail to work hard or well enough
- It cannot stop employees leaving the firm

binding an employee to a job for a long time is illegal. And because employees sometimes do leave, the firm would like to have a large pool of qualified applicants in line as replacements.

An employment contract omits things that both the employees and the business owner care about: how hard and well the employee will work, and for how long the worker will stay. As a result of this *contractual incompleteness*, paying the least possible wage is almost never the firm's strategy to minimise the cost of acquiring the labour effort it needs.

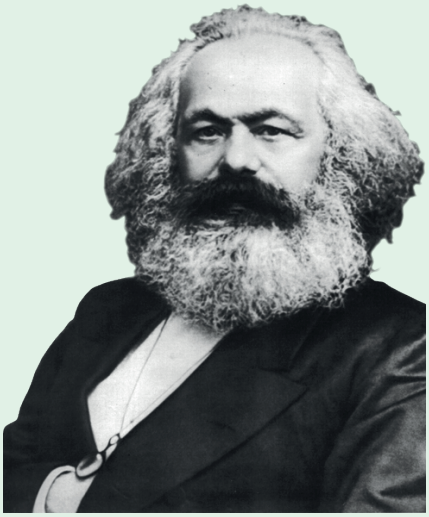
DISCUSS 6.2: INCOMPLETE CONTRACTS

Think about some of the more common jobs you see around you: perhaps a teacher, a banker, a retail worker, a nurse or a police officer.

Which of these are the easiest to write a contract for, and why?

GREAT ECONOMISTS

KARL MARX



Adam Smith, writing at the birth of capitalism in the 18th century, was to become its most famous advocate. Karl Marx (1818-1883), who watched capitalism mature in the industrial towns of England, was to become its most famous critic.

Born in Prussia (now part of Germany), he distinguished himself as a student at a Jesuit high school only by his rebelliousness. In 1842 he became a writer and editor for the *Rheinische Zeitung*, a liberal newspaper, which was then closed by the government, after which he moved to Paris, met Friedrich Engels, with whom he collaborated in writing *The Communist Manifesto* (1848), and moved to London in 1849. At first Marx and his wife Jenny lived in poverty. He earned money by writing about political events in Europe for the *New York Tribune*.

Marx saw capitalism as just the latest in a succession of economic arrangements in which people have lived since prehistory. Inequality was not unique to capitalism, he observed—slavery, feudalism, and most other economic systems had shared this feature—but capitalism also generated perpetual change and growth in output.

He was the first economist to understand why the capitalist economy was the most dynamic in human history. Perpetual change arose, Marx observed, because capitalists could survive only by introducing new technologies and products, finding ways of lowering costs, and by reinvesting their profits into businesses that would perpetually grow.

This, he claimed, inevitably caused conflict between employers and workers. Buying and selling goods in a market appeared to be transactions among equals: nobody is in a position to order anyone else to buy or sell. In the labour market, in which owners of capital are buyers and workers are the sellers, the appearance of freedom and equality was, to Marx, an illusion.

Employers did not buy the employee's work, because this cannot be purchased, as we have seen in this unit. Instead the wage allowed the employer to rent the worker and to command workers inside the firm. Workers were not inclined to disobey because they might lose their job and join, in the phrase that Marx used in *Capital* (1867), the "reserve army" of the unemployed. Marx thought that the power wielded by employers over workers was a core defect of capitalism.

Marx also had influential views on history, politics, and sociology. He thought that history was decisively shaped by scarcity and technological progress interacting with economic institutions, and that political conflicts arose from conflicts about the distribution of income and the organisation of these institutions. He thought that capitalism, by organising production and allocation in anonymous markets, created atomised individuals instead of integrated communities.

In recent years economists have returned to themes in Marx's work to help explain economic crises. Among these themes: the firm as an arena of conflict and of the exercise of power (this unit), the role of technological progress (Unit 1 and Unit 2), and the problems created by inequality (Unit 19).

Why is it not possible for firms just to pay employees according to how productive they are? For example, paying employees at a clothing factory $2 for each garment they finish. This method of payment, known as *piece rate*, provides the employee with an incentive to exert effort; employees take home more pay if they make more garments.

In the late 19th century the pay of more than half of US manufacturing workers was based on their output, but piece rates are not widely used in modern economies. At the turn of the 21st century less than 5% of manufacturing workers in the US were paid piece rates and, beyond the manufacturing sector, piece rates are used even less often.

Why do today's firms not typically use this simple method to induce high effort from their employees?

- It is *very difficult to measure* the amount of output an employee is producing in modern knowledge- and service-based economies (think about an office worker, or someone providing home care for an elderly person).
- *Employees rarely work alone*, so measuring the contribution of individual workers is difficult (think about a team in a marketing company working on an advertising campaign, or the kitchen staff at a restaurant).

If piece rates are not practical, then what other method could a firm use to induce high effort from workers? How could the firm provide an incentive to do the job well, even though the worker is paid for time and not output? Just as the owners of the firm protect their interests by linking management pay to the firm's share price, the manager uses incentives so that employees will work effectively.

6.4 EMPLOYMENT RENTS

There are many reasons why people put in a good day's work. For many people, doing a good job is its own reward; doing anything else would contradict our work ethic. Even for those not intrinsically motivated to work hard and well, feelings of responsibility for other employees, or for one's employer, may provide strong work motivations.

For others, hard work is a way that the employee can reciprocate a feeling of gratitude to the employer for providing a job with good working conditions. In other cases firms identify teams of workers whose output is readily measured—the percentage of on-time departures in the case of airport staff of airlines for example—and pay a benefit to the whole group that is divided among team members.

But, in the background, there is another reason to do a good job: fear of being fired, or of missing the opportunity to be promoted into a position that has higher pay and greater security against being laid off. A slacking employee has good reason to fear losing a job:

- *Owners' right to fire*: The owners of the firm, by definition, have the right to exclude the worker—that is, to terminate employment.
- *Employment rents*: Rents make this a meaningful threat. Employees are typically paid much more than the minimum they would accept for taking the job, so they are receiving an economic rent (called an employment rent), meaning that they would prefer to keep the job rather than ending up with the reservation option.

Because owners and managers decide whether an employee stays or goes, they can benefit from the implicit threat of the sack to make the worker perform in ways that that person would not choose unless the threat was real. This means that the owners and managers exert power over employees. The employment rent is a measure of what we call the cost of job loss.

WHEN ECONOMISTS *AGREE*

RONALD COASE AND KARL MARX ON THE FIRM AND ITS EMPLOYEES

George Bernard Shaw (1856-1950), a writer, joked that "If all economists were laid end to end, they would not reach a conclusion."

This is funny, but not entirely true.

For example, the two leading economists of the early 19th century—Ricardo and Malthus—were political opponents. Ricardo often sided with businesspeople, for example in supporting freer imports of grain to Britain so as to reduce food prices and allow lower wages. Malthus opposed him and supported the Corn Laws that restricted grain imports, a position favoured by the landed gentry. But the two economists independently developed the same theory of land rents, which we still use today.

Even more striking is that two economists from different centuries and political orientations came up with similar ways of understanding the firm and its employees.

In the 19th century Marx contrasted the way buyers and sellers interact on a market, voluntarily engaging in trade, with how the firm is organised as a top-down structure, one in which employers issue orders and workers follow them. He called markets "a very Eden of the innate rights of man", but described firms as "exploit[ing] labour-power to the greatest possible extent."

When Ronald Coase (an economist, who we will study in more detail in Unit 10) died in 2013, he was described by Forbes magazine as "the greatest of the many great University of Chicago economists". The motto of Forbes is "The capitalist tool", and the University of Chicago has a reputation as the centre of conservative economic thinking.

Yet, like Marx, Coase stressed the central role of authority in the firm's contractual relations:

"Note the character of the contract into which an [employee] enters that is employed within a firm... for certain remuneration [the employee] agrees to obey the directions of the entrepreneur."
— Ronald H. Coase, *The nature of the firm* (1937)

Recall that Coase had also defined the firm by its political structure: "If a workman moves from department Y to department X, he does not go because of a change in prices but because he is ordered to do so." He sought to understand why firms exist at all, calling them "islands of conscious power in this ocean of unconscious cooperation."

Both based their thinking on careful empirical observation, and they arrived at a similar understanding of the hierarchy of the firm. They disagreed, however, on the consequences of what they observed: Coase thought that the hierarchy of the firm was a cost-reducing way to do business. Marx thought that the coercive authority of the boss over the worker limited the employee's freedom. Like Malthus and Ricardo, Coase and Marx disagreed. But like the Malthus and Ricardo, they also advanced economics with a common idea.

We can use the same reasoning in the employment of managers by the owners of the firm. The main reason owners wield power over managers is that they can fire them, and so eliminate their managerial employment rents.

Recall that a *rent* measures the value of a situation—having your current job, for example—compared to what you would get if the current situation were no longer possible.

The *cost of job loss* (also termed the *employment rent*) for a worker includes:

- *Lost income* while searching for a new job (perhaps partially offset by an unemployment insurance benefit or, in poorer countries, by the possibility of lower-paying self-employment or work on the family farm).
- *Loss of firm-specific assets* such as the psychological costs of losing workplace friends, and possibly needing to relocate with one's family to some other locality where jobs are easier to get.
- *Loss of medical insurance* available through an employer in some countries.
- *The social stigma of being unemployed*, which as we will see in Unit 12, for most people is equivalent to a substantial financial cost.

Even confining attention to the loss in wages, the cost is high. But how do we measure how high it is?

HOW ECONOMISTS LEARN FROM FACTS

HOW LARGE ARE EMPLOYMENT RENTS?

Set aside the undoubtedly large, but hard-to-measure, psychological and social cost of losing one's job, estimating the cost of job loss (the size of the employment rent) is not simple.

Can we compare the economic situation of workers currently employed with the economic situation of unemployed people? No, because the unemployed are different people, with different abilities and skills. Even if they were employed, they would be likely to earn less than people who currently have jobs.

An entire firm closing, or a mass layoff of workers, provides a *natural experiment* that can help. We could look at the earnings of workers before and after they lost their job during a major employment cutback. When a factory closes because the parent company has decided to relocate production to some other part of the world, for example, virtually all workers lose their jobs not just the ones who were most likely to lose their jobs through poor performance.

Louis Jacobson, Robert Lalonde and Daniel Sullivan used just such a natural experiment to estimate the cost of job loss. They studied experienced (not recently hired) full-time workers hit by mass layoffs in the US state of Pennsylvania in 1982. In 2014 dollars, those displaced had been averaging $55,000 in earnings in the year before they lost their jobs. Those who were fortunate enough to find another job in less than three months took jobs that paid a lot less, averaging only $35,000: they lost $20,000 in the first year after they were laid off.

Four years later they were still making $12,000 less than similar workers who had been making the same initial wage, but whose firms did not lay off their workers. In the five years that followed their layoff they lost the equivalent of an entire year's earnings.

Many, of course, did not find work at all. They suffered even greater costs.

The year 1982 was not a good time to be looking for work in Pennsylvania, but similar estimates (from the US state of Connecticut between 1992 and 2004 for example) suggest that even in better times employment rents are large enough that workers would worry about losing them.

6.5 DETERMINANTS OF THE EMPLOYMENT RENT

To understand what determines the size of the employment rent, think about Maria's situation. We do not compare Maria with someone who is out of work, but with Maria herself *if she were to lose her job*.

To do this we need to think how Maria would evaluate two aspects of her job:

- *The pay that she gets*: That is, something she values.
- *How hard she works*: She would like to do no more work than is necessary.

We can compare these two dimensions using the concept of *utility* that we introduced in Unit 3. The utility of the goods and services Maria can buy with her wage is the satisfaction she derives from having them. Similarly the unpleasantness of going to work and working hard all day is a source of dissatisfaction, that we call her *disutility*. In the model she evaluates the whole package—the utility from the wage and the disutility of the work. Use the slideline in Figure 6.2 to find out how to calculate Maria's employment rent in this case. Remember that in this example Maria's wage is $12; at this wage she spends half of her working time actually working, and half doing other things like checking Facebook.

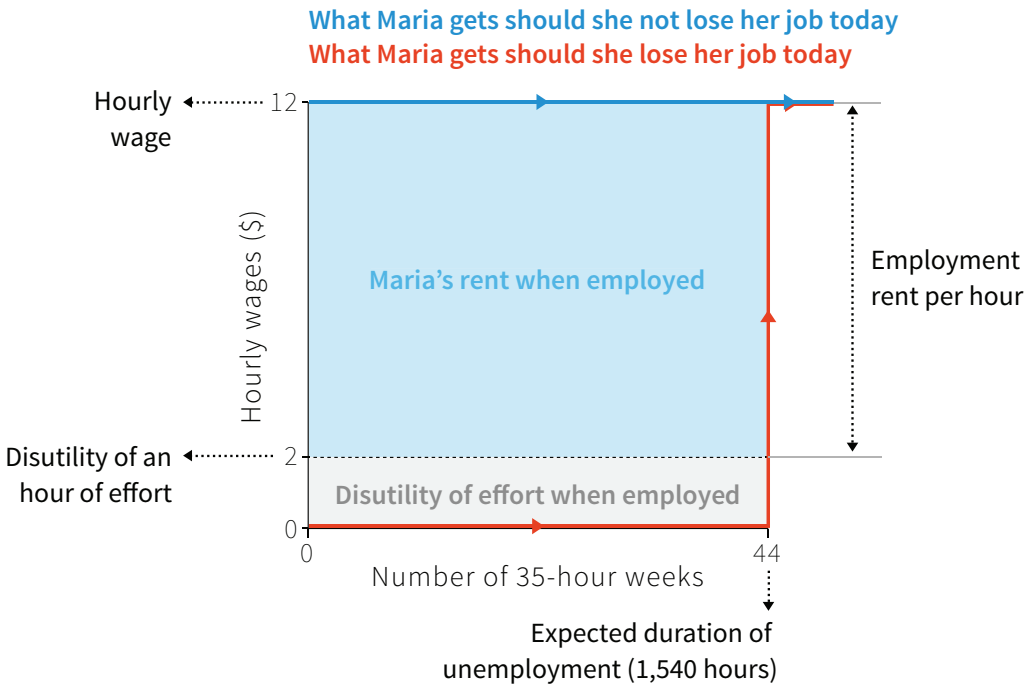


Figure 6.2 *Maria's employment rent for a given effort and $12 wage in an economy without an unemployment benefit.*

Suppose that Maria receives an hourly wage that, after the payment of taxes and other deductions, is $12. She will continue to receive this for the foreseeable future if she keeps her job, indicated by the horizontal line in the figure. Maria's employer would like her to work harder than she would prefer. When paid $12, half of the time on the job she pursues non-work activities, which we represent as an effort level of 0.5. Working this hard is equivalent to a cost of $2 per hour to Maria. The difference between her wage per hour and her disutility of effort per hour is the *net benefit* per hour that she receives from being employed. If Maria were to lose her job at time 0, she would no longer receive her wages; and this unfortunate state would persist as long as her spell of unemployment goes on, indicated by the horizontal line at the bottom of the figure. The expected duration of her unemployment is simply the number of 35 hour weeks that she will remain without pay (and also without the disutility of working). Maria finds a job at the same wage as before after being unemployed for 44 weeks (or *44 x 35 = 1540 hours*). The large shaded area is her *total cost of job loss* from the spell of unemployment, that is, her employment rent.

So, to calculate Maria's employment rent we need to compare what she is getting on the job with what she would get were she to lose the job:

- What she gets on the job is her wage.
- What she would get, if she lost her job, is the satisfaction of not having to work, which she values at $2 per hour.

Using the data from Figure 6.2, Maria's cost of job loss per hour on the job is:

$$\begin{aligned} \textit{employment rent per hour} &= \textit{wage} - \textit{disutility of effort per hour} \\ &= \$10 \end{aligned}$$

Her total employment rent is the employment rent per hour, times the number of hours of work she will lose if her job is terminated. It is the shaded area in the last panel of the figure.

$$\begin{aligned} \textit{employment rent} &= \textit{employment rent per hour} \times \textit{expected lost hours of work} \\ &= \$10 \textit{ per hour} \times 1{,}540 \textit{ hours} \\ &= \$15{,}400 \end{aligned}$$

People who lose their jobs typically can expect help from family and friends while they are out of work. Also, in many economies, people who lose their jobs may receive *unemployment insurance from the government* or financial assistance. In poorer economies, they may be able to earn a small amount in informal self-employment.

Thus we can add what we will call an unemployment benefit (including assistance from nongovernmental sources and self-employment income) as a partial offset to Maria's lost wage income. But we assume that Maria's unemployment benefit runs out eventually: families and friends will not be able to help for ever, and government unemployment insurance is often time-limited.

Figure 6.3 shows Maria's employment rent as it would be if she were eligible for an unemployment benefit, and if she continues working at the same pace, so that her disutility of effort is unchanged.

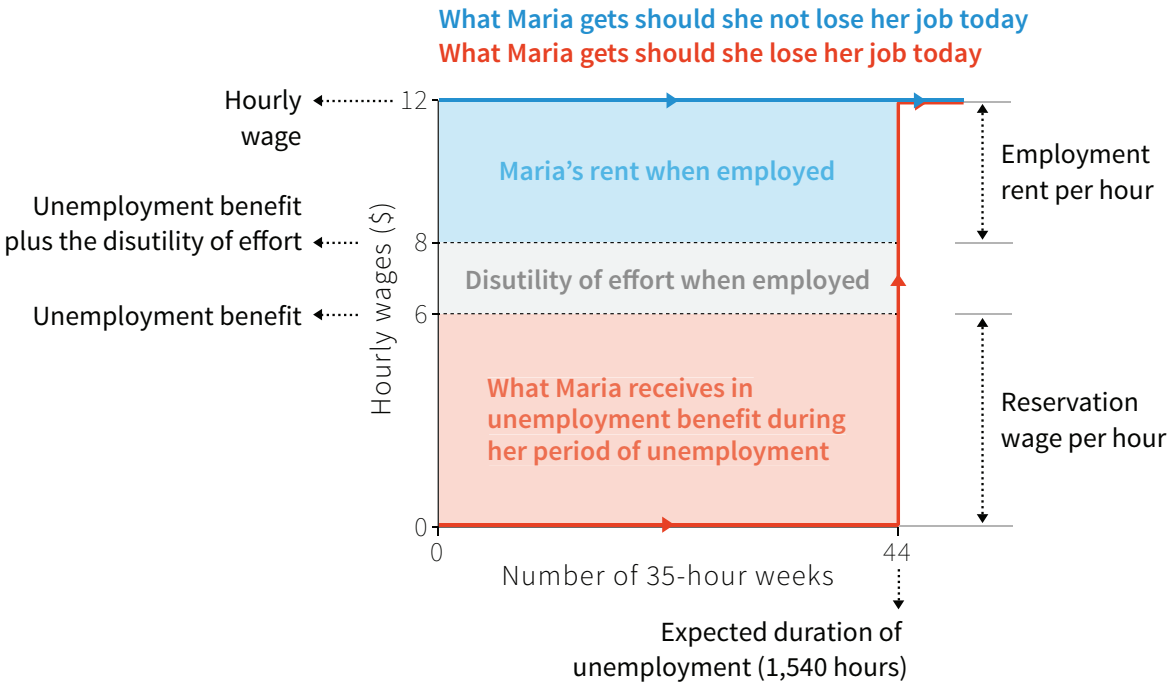


Figure 6.3 *Maria's employment rent for a given effort and a $12 wage in an economy with an unemployment benefit.*

The $6 in unemployment benefit is called Maria's *reservation wage*. In Figure 6.2, without unemployment benefits, her reservation wage was zero. If the wage remains $12, from the data in the figure and the above reasoning we now see that:

$$\begin{aligned} \textit{employment rent} &= \textit{wage} - \textit{reservation wage} - \textit{disutility of effort} \\ &= \textit{wage} - \textit{unemployment benefit} - \textit{disutility of effort} \\ &= \$12 - \$6 - \$2 \\ &= \$4 \end{aligned}$$

And taking account of the duration of unemployment we see that:

$$\begin{aligned} \textit{employment rent} &= \textit{employment rent per hour} \times \textit{expected hours of lost work time} \\ &= \$4 \textit{ per hour} \times 1{,}540 \textit{ hours} \\ &= \$6{,}160 \end{aligned}$$

DISCUSS 6.3: ASSUMPTIONS OF THE MODEL

As in all economic models, our simplified representation of Maria's employment rent has deliberately not considered some aspects of the problem that might be important to explore, depending on what we want to know. For example, we have assumed that:

1. Maria finds a job with the same pay after her spell of unemployment.
2. Maria benefits from not expending effort when unemployed (the lack of the disutility of effort), but does not bear any of the psychological or social costs of joblessness.

For both, redraw Figure 6.3 to show how taking them into account would alter the figure.

Our next step is to use the concepts of employment rent and reservation wage to see how employers set wages, and how employees respond with a level of work effort. To study the social interaction between the employer and the employee we need to examine how the employer sets the wage when:

- The employer knows that the wage affects the worker's employment rent.
- The employment rent will influence how hard the employee works.

6.6 WORK AND WAGES: THE LABOUR DISCIPLINE MODEL

We can now create a model of social interaction in the firm: we represent the way the owners of the firm, through their managers, interact with employees as a game. Remember that a game is a description of a social interaction including a list of the players, the strategies they can adopt, the order in which the players choose their actions and what they know when they do, and the outcomes for each of the players (their payoffs) for all of the strategies that may be chosen. As with other models, we ignore some aspects of their interactions to focus on what is important, following our motto of "seeing more by looking at less".

On the stage of the firm the cast of characters is now just the owner (the employer) and a single worker, Maria. The game that describes their interactions is sequential (one of them chooses first, like the ultimatum game) and the game is repeated. Here is the order of play:

1. *The employer chooses a wage*, based on his knowledge of how the employee will respond to higher or lower wages, and informs Maria that she will be employed in subsequent periods at the same wage—as long as she works hard enough to satisfy him.
2. *Maria then chooses a level of work effort* that is her best response to the wage offered, and to the prospect of losing her job if she does not provide enough effort.

The employer's problem is this:

- Maria's effort is essential to producing the goods or services necessary for him to make a profit.
- The greater her effort, the greater will be the output and hence, *ceteris paribus*, his profits.
- He will maximise his profits by acquiring a given amount of Maria's effort at the least possible cost. This is equivalent to acquiring the most possible effort and therefore output and profit for a given wage cost.

Maria's best response to the wage offered will be the effort level that balances her desire to keep her job with her desire also to not exhaust herself by working.

Then, taken together, the wage rate chosen by the employer and the effort level chosen by Maria are a Nash equilibrium. This allocation represents the best that each can do given:

- The way that Maria responds to the wages her employer might offer
- The wage that he chose

The payoff for the employer is the profit; Maria's payoff is her valuation of the wage she receives, taking into account the effort she has expended.

Employers typically hire work supervisors and install surveillance equipment to keep watch on their employees, increasing the likelihood that the management will find out if a worker is not working hard and well. Here we will ignore these extra costs and just assume that the employer occasionally gets some information on how hard or well an employee is working. This is not enough to implement a piece-rate contract, but more than enough to fire a worker if the news is not good. Maria knows that the chance of the employer getting bad news decreases the harder she works.

The employee's best response

When the cost of job loss (the employment rent) is large, workers will be willing to work harder. Holding constant other ways that it might influence the employment rent, an employer can increase the cost of job loss by raising wages.

To decide on the wage to set, the employer needs to know how the employee's work effort will respond to higher wages. In the example in Figure 6.3, Maria was paid $12 per hour and had an effort level of 0.5. Figure 6.4 shows the resulting relationship between effort and wages, referred to as the worker's *best response curve*. (It can also be called a best response *function* because a function gives the value of one variable, in this case the worker's effort level, given some values of another variable, in this case, the wage.)

Effort per hour, measured on the vertical axis, varies between zero and one. We can think of this as the proportion of each hour that Maria spends working diligently (the rest of the time she is not working). For example, an effort level of 0.5 indicates Maria is spending half the day on non-work related activities such as checking Facebook, shopping online, or just staring out of the window.

Think of the best response curve as the answer to a hypothetical "if-then?" question. It gives the answer to: "If the wage is $12 then what effort will the worker put in?" as well as answers to the identical questions for all other possible wages that the employer might offer.

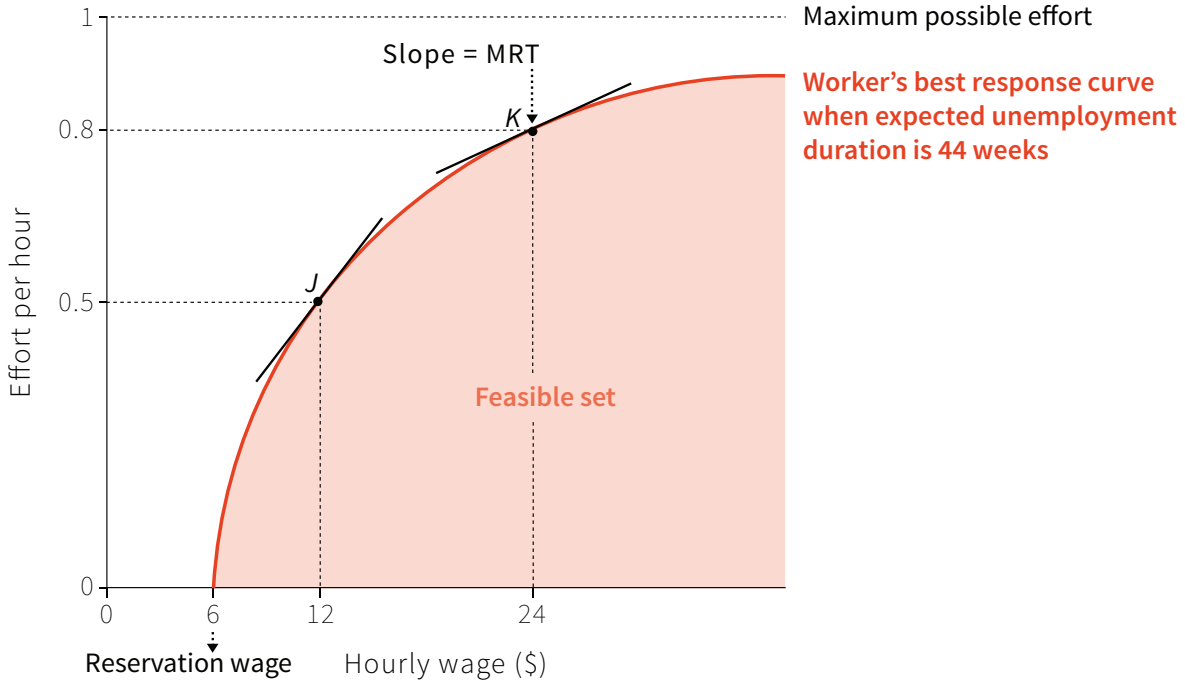


Figure 6.4 *Maria's best response to the wage. Point J refers to the information in Figure 6.3.*

Maria's best response curve is similar to Angela's production function for grain in Unit 3, which showed how giving up free time translated into more goods. It is also similar to Alexei's function translating study time into grades. Maria's best response curve shows how a cost to the employer (higher wages) translates into something the employer values: more effort, and therefore more output, from his workers.

The best response curve also becomes flatter as the wage and the effort level increase. This is because, as the level of effort approaches the maximum possible, the disutility of effort becomes greater. In this case it takes a larger employment rent (and hence a larger wage) to get effort from the employee.

Just like the production functions of Alexei and Angela in Unit 3, an employer who pays the worker a higher wage faces diminishing marginal returns. In other words, the higher the initial wage, the smaller the increase in effort and output the employer gets from a $1 per hour increase in wages.

Seen from the standpoint of the owner or the employer, the slope of the best response curve is the *marginal rate of transformation* of paying higher wages into getting more work from the worker. The best response curve is therefore the frontier of the employer's feasible set of combinations of wages and effort that it gets from its employees.

To discover the properties of a worker's best response function using calculus, see the Leibniz.

The reservation wage and the employee's best response curve

What is the lowest wage that will motivate the employee to provide any on the job effort at all? To answer this, think about the wage that is so low that she doesn't care if she is fired, so there is no motivation to exert effort.

Going back to Maria, imagine that in the case depicted in Figure 6.3 she were offered a wage of $6, identical to what she would get per hour from her unemployment benefit over the time she would be unemployed. Even if she put in no work whatsoever (and so endured no disutility of effort) her job at a $6 wage would be no better than being without work. So she would not care one way or the other if her job ended. Thus, as you can see in the figure, the reservation wage is the wage at which the best response curve hits the horizontal axis.

As a result if the firm were for some reason to pay the reservation wage, the employee would not work. This is why the firm will not offer the lowest wage possible.

6.7 WAGES, EFFORT AND PROFITS IN THE LABOUR DISCIPLINE MODEL

Remember Maria is not in the situation that Angela faced when Bruno could order her to work at the point of a gun. Maria has bargaining power because she can always walk away, an option that, initially, Angela did not have.

Maria chooses how hard she works. The best the owner can do is to determine the conditions in which she makes that choice. The owners and managers know that they cannot get Maria to provide more effort than is given by the best response curve shown in Figure 6.4. The fact that the best response curve slopes upwards means that employers face a trade-off: *they can get more effort only by paying higher wages*.

The employer can choose any combination of wage and effort on (or to the right of) the best response curve. To decide on the wage to set the employer thinks about the Maria's effort just as he thinks about any other input: he wants to minimise its cost. When the company is purchasing an input, say a chemical used in the production process, the employer finds the supplier that provides the greatest quantity for a given expenditure, that is, he looks for the lowest price at which the chemical can be acquired.

In the same way, he is looking for a way to maximise the amount of effort Maria provides for the wages paid. But, unlike the case of purchasing chemicals (finding the lowest price is the objective) in the case of labour, the employer is not trying to pay the lowest possible wage. The employee's best response curve tells him that if he paid the reservation wage, the workers might show up (they wouldn't care one way or the other), but they would not work if they did.

So for a given level of the other influences on profits (revenues from sales, how much output each unit of effort produces, and other costs), the employer wants the cost of effort to be as low as it can be. In order to maximise profits under these *ceteris paribus* assumptions, the owner finds the wage such that effort per dollar spent on wages will be as high as it can be, given the employee's best response function.

So, to maximise profits, the employer maximises the ratio of effort to wages or, using e for effort and w for wages, he maximises e/w. Since effort by workers is necessary to produce output, he tries to minimise the cost of producing a unit of output, including the cost of effort and other inputs. A line like the one shown in Figure 6.5 has a slope of e/w. We call this line an *isoprofit curve* because, holding constant the other influences on profits, the employer's profits per unit of output produced are the same for every point on the curve.

Use the slidebar in Figure 6.5 to see that all the points on the diagonal line have the same effort-wage ratio. Points on the line, like $e = 0.45$ and $w = \$10$ have the same e/w ratio as $e = 0.9$ and $w = \$20$, namely 0.045. As far as the employer is concerned, if the firm could get $e = 0.9$ by paying \$20 and $e = 0.45$ by paying half that amount; he would be equally happy, as profits per unit of output would be the same at either of these points. Assuming that his only concern is profits, this line is an indifference curve, just like the indifference curves already studied.

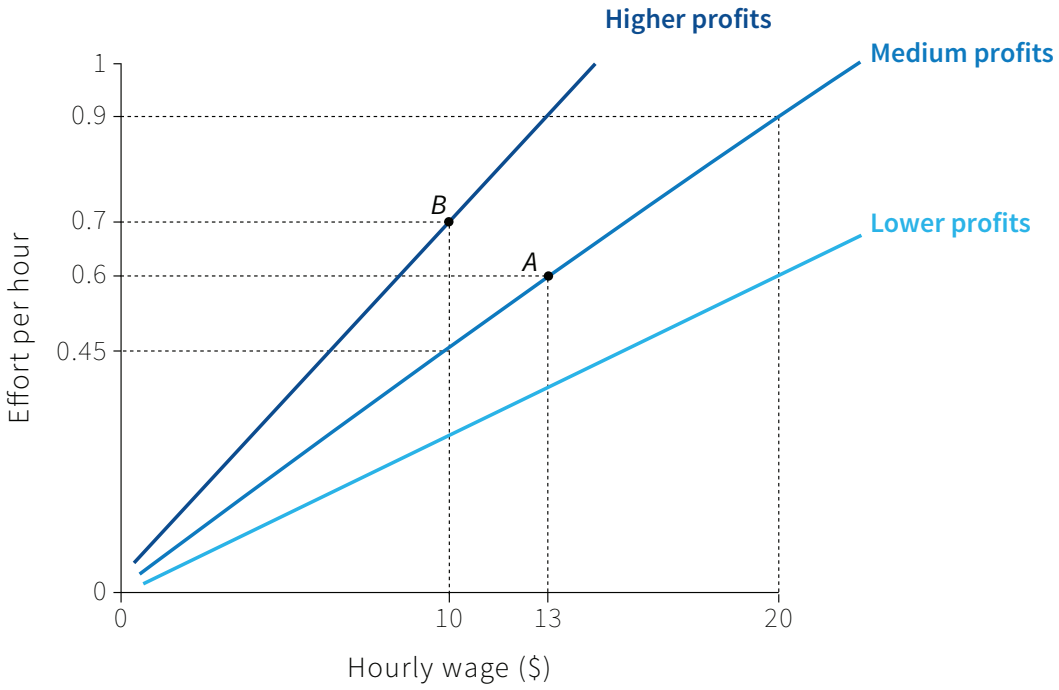


Figure 6.5 *The employer's objectives: the isoprofit curves.*

Like other indifference curves, the slope of the isoprofit curve is the marginal rate of substitution of wages for effort levels. The curve slopes upwards because starting from any point on it, a higher effort level must be accompanied by a higher wage for the effort/wage ratio to remain unchanged. Consider these two additional isoprofit lines. The slope of each line is the e/w ratio. A steeper line means a lower cost of effort and higher profits for the employer. On the steepest isoprofit curve an effort level of 0.7 costs only \$10 in wages (B), whereas on the flatter middle curve, effort at this wage is only 0.45. Some lines are better for the owners than others.

To maximise profits, the employer will seek to get onto the highest isoprofit line possible. But because he cannot dictate the level of effort, he has to pick some point on Maria's best response curve.

The best he can do is to set the wage at \$12 on the isoprofit line that is tangent to Maria's best response curve (point A). Use the slidebar in Figure 6.6 to see how the employer sets the wage.

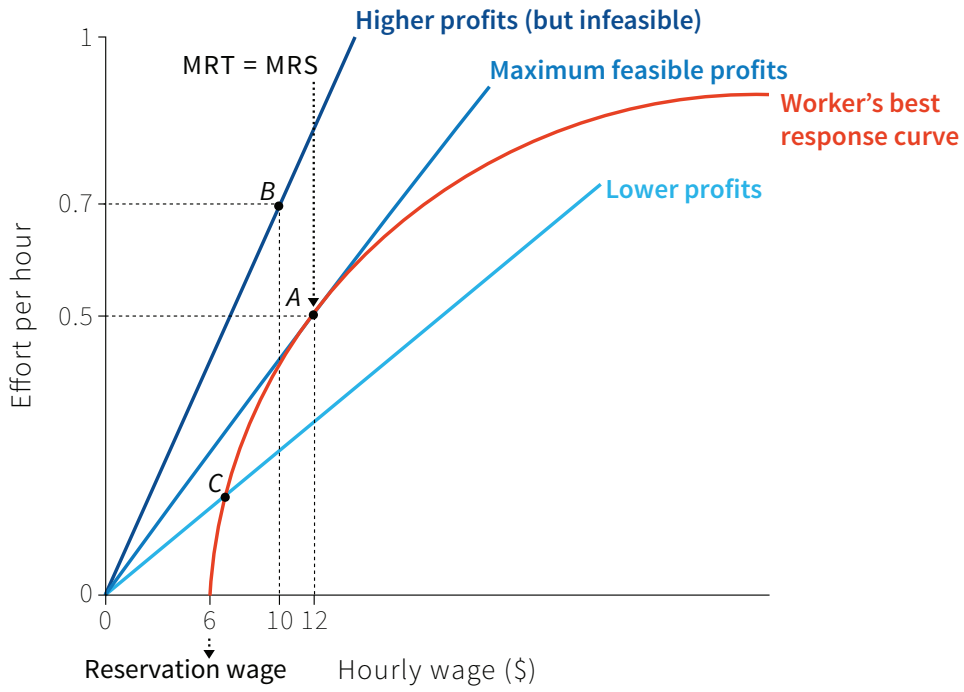


Figure 6.6 *The employer sets the wage to maximise profits*

The firm will choose to pay a higher wage than at *C* to benefit from a higher effort/wage ratio. Profits are maximised at *A* where MRS = MRT.

In Figure 6.6, the employer will choose point *A*, offering a wage of $12 per hour to hire Maria, who will exert effort of 0.5. The employer cannot do better than this point: for example, point *B* would provide better profits, but it is infeasible.

This is similar to individuals choosing a level of work to maximise their utility in Unit 3 and Unit 5. There the aim was to get on to the highest feasible indifference curve by finding the point at which an indifference curve was tangent to the feasible consumption frontier, and working at the level indicated by that point. In Unit 3 the tangency occurred where the marginal rate of substitution (the slope of the indifference curve) equalled the marginal rate of transformation (the slope of the feasible consumption frontier).

The same is true in this case. The marginal rate of substitution on the indifference curve, which is the slope of the isoprofit line, is equal to the marginal rate of transformation of higher wages into greater effort.

The Leibniz shows you how to find the profit-maximising wage for an employer, using calculus.

When wages are set by the employer in this manner they are sometimes called *efficiency wages* because the employer is recognising that what matters for profits is e/w, the units of effort (called *efficiency units*) per dollar of wage costs, rather than how much an hour of work costs.

What has the *labour discipline model* told us?

- *Equilibrium*: In the owner-employee game, the employer offers a wage and Maria provides a level of effort in response. It is a Nash equilibrium allocation (outcome).
- *Rent*: In this allocation Maria provides effort because she receives an employment rent that she might lose were she to slack off on the job.
- *Power*: Because Maria fears losing the rent, the employer is able to exercise power over her, getting her to act in ways that she would not do in the absence of the threat of job loss. This contributes to the profits of the employer.
- *Involuntary unemployment*: The fact that Maria is receiving a rent means that were she to lose her job, she would be worse off. She could only not be worse off if she could get re-employed immediately at the same wage. She would therefore be involuntarily unemployed. Just as she prefers to keep her job rather than being unemployed, there are other people who would prefer to have her job rather than remaining unemployed; so there is unemployment in the equilibrium of the game.

The chain of reasoning that leads to unemployment can be expressed in a different way, which is called a proof by contradiction.

Suppose there is no involuntary unemployment, then:

- The duration of unemployment must be zero, meaning...
- ... employment rents are zero, so...
- ... no work is done and no output produced, in which case...
- ... employers have no reason to employ anyone, so that...
- ... nobody is employed and everyone is unemployed.

This contradicts the initial statement that there is no unemployment. So the initial assumption, "suppose that there is no involuntary unemployment" cannot be true.

Unemployment is an important concern for voters and the policymakers who represent them. A key feature of the model is that there is involuntary unemployment. We use the term involuntary unemployment because the unemployed are not jobless by choice: they are identical to those with jobs and would prefer to have work at the same terms as those with jobs are receiving.

The model also shows how the policies that governments pursue to alter the level of unemployment, or to provide income to unemployed workers, will affect the profits of firms and the effort level of their employees.

DISCUSS 6.4: THE EMPLOYER SETS THE WAGE

Would either of the following affect the curves in Figure 6.6? If so, explain in what way the curves are affected.

1. The government decides to increase the provision of subsidised child care.
2. Demand for the firm's output rises as celebrities endorse the good.

6.8 PUTTING THE MODEL TO WORK: OWNERS, EMPLOYEES AND THE ECONOMY

Until now we have discussed how the employer will select some point on the best response function. But public policies can shift the entire best response function, moving it to the right (or up) or to the left (or down). To see how this works, recall that the position of the best response function depends on:

- The importance to the employee of the things that can be bought with the wage
- How unpleasant it is to exert effort
- The probability of getting fired when working at each effort level
- The worker's reservation wage

If there are changes in any of them, the best response curve will either shift to the right or to the left.

First, imagine how a change in the duration of unemployment will affect the position of the best response curve. Recall that the unemployment benefit including support from family and friends is limited, so the longer the spell of unemployment, the lower the level of the unemployment benefit per hour of lost work (or per week). So an increase in the duration of a spell of unemployment has two effects:

- It reduces the reservation wage, increasing the employment rent.
- It extends the period of lost work time, and hence lost employment rents should she lose her job.

Figure 6.7 shows how, as a result, the best response curve shifts left. The new curve means that, for any given wage, the worker will supply more effort than before, improving the profit-making conditions for the employer.

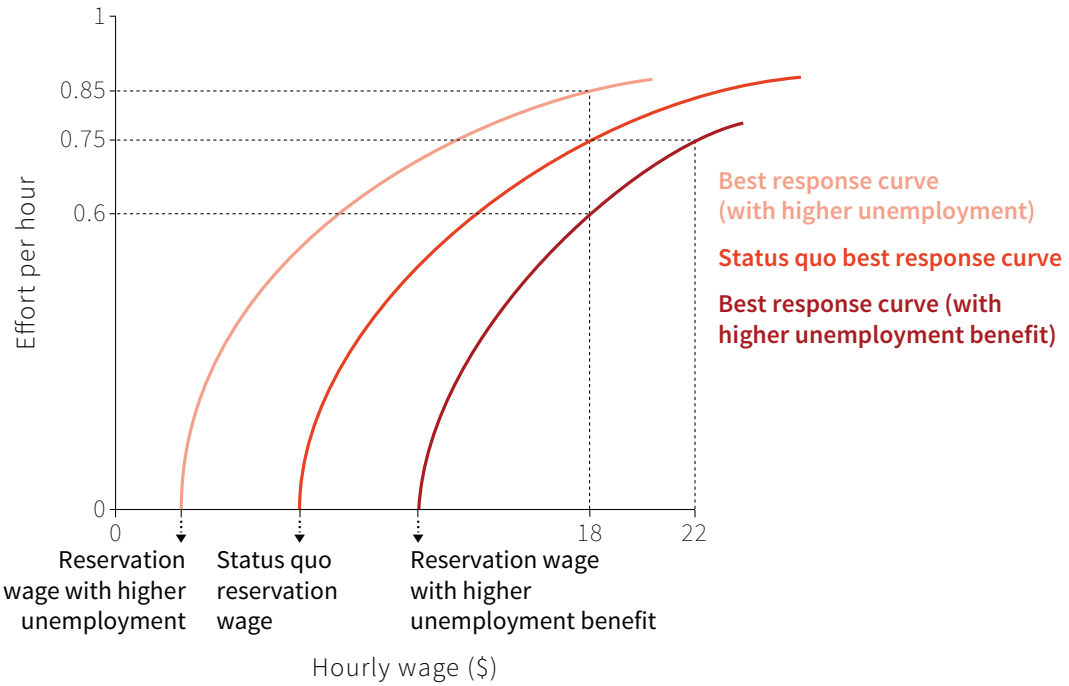


Figure 6.7 *The best response curve depends on the level of unemployment and the unemployment benefit.*

As shown in Figure 6.7 an increase in the level of unemployment benefit raises the reservation wage and, as a result, for any given wage it reduces the amount of effort the worker will provide. A rise in the unemployment benefit thus shifts the worker's best response curve to the right.

If we choose a level of effort, say 0.6, and ask how much the employer would have to pay to get the worker to provide that amount of work, we see that the wage is lower if unemployment increases from the initial situation, and higher if the unemployment benefit increases. From this point of view the curve is shifting to the left or right. Both descriptions are correct; it is just the point of view that differs.

DISCUSS 6.5: EFFORT AND WAGES

An employer faces the best response curves shown in Figure 7. The employer would like to elicit a level of effort of 0.75. In each of these three cases, explain why the wage differs.

DISCUSS 6.6: DURATION OF UNEMPLOYMENT

To this point, we have used figures with effort on the vertical axis and hourly wage on the horizontal axis. If instead we chose a given wage, say $12 an hour, and let the expected duration of unemployment vary, putting it on the horizontal axis instead of the wage:

1. What would a worker's best response curve look like?
2. What would happen to this curve if the given wage increased to $14?
3. What if unemployment benefit were raised?

Economic policies can alter both the size of the unemployment benefit and the extent of unemployment (and hence the duration of a spell of unemployment). These policies are often controversial because they shift the employee's best response function either to the right (favouring employees, who then will work less hard for any given wage) or to the left (favouring owners who will as a result acquire the effort of their employees at less cost, raising profits).

We will now see that shifts in the employee's best response function can alter the functioning of the entire economy by affecting the feasible combinations of real wages and employment.

6.9 THE FIRM AND ITS EMPLOYEES IN THE ECONOMY

We can now broaden the perspective from a single firm to the economy as a whole. We ask how changes in the unemployment rate affect the wage set by employers.

In Figure 6.8 the employment rate in the whole economy is on the horizontal axis. You will see that it goes up to a value of 1.

- The *employment rate*: defined as the proportion of people of working age, usually defined as those between 16 and 64, who are employed.
- The *labour force*: The vertical line at a value of the employment rate less than one.
- *Inactive workers*: Between the labour force line and the employment rate of one is the proportion of people of working age who are neither working nor actively looking for work; they are out of the labour force.

- The *unemployment rate*: The proportion of those in the labour force who are not employed.

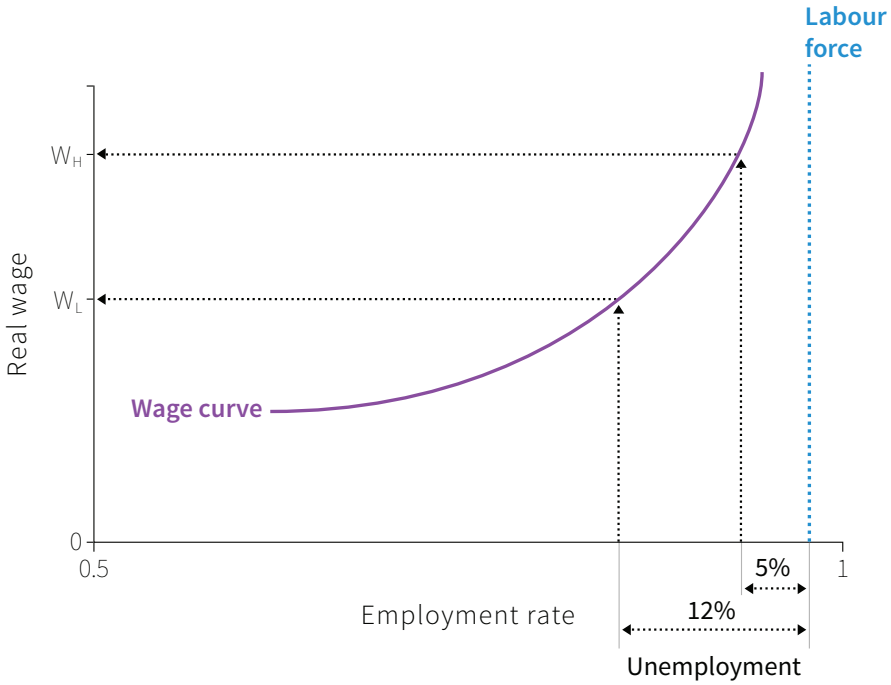


Figure 6.8 *The wage curve: labour discipline and unemployment in the economy as a whole.*

The upward-sloping line is called the *wage curve*. At 12% unemployment in the economy, the employee's reservation wage is low and the worker will put in a high level of effort for a relatively low wage. The firm's profit-maximising wage is therefore low. At 5% unemployment in the economy, the employee's reservation wage is high and they will not put in much effort unless the wage is high. The firm's profit-maximising wage is therefore higher.

The upward-sloping line is called the *wage curve*. Like the effort best response function of the employee on which it is based, the wage curve is a mathematical version of an "if-then" statement: if the employment rate is *X* the Nash equilibrium wage will be *Y*. This means that at the employment rate *X* the wage *Y* is the result of both employers and employees doing the best they can in respectively setting wages and responding to the wage with a given amount of effort.

This statement is true because the wage curve for the whole economy is based directly on the employer's wage-setting decision and the employee's effort decision in an economy that is composed of a great many firms like the firm we have just modelled.

We show how to do this in Figure 6.9 by bringing together Figure 6.8 (the economy-wide wage curve) and Figure 6.6 (how the firm sets the wage).

In the top panel of Figure 6.9 is the employee's best response curve at the two unemployment rates of 12% and 5%. Recall that a higher unemployment rate reduces the reservation wage, because a worker faces a longer expected period of unemployment if he or she loses a job. This weakens the bargaining power of the employee and shifts the best response curve to the left: with an unemployment rate of 12%, the reservation wage is shown by point *F*. The employer's profit-maximising choice is point *A* with the low wage w_L.

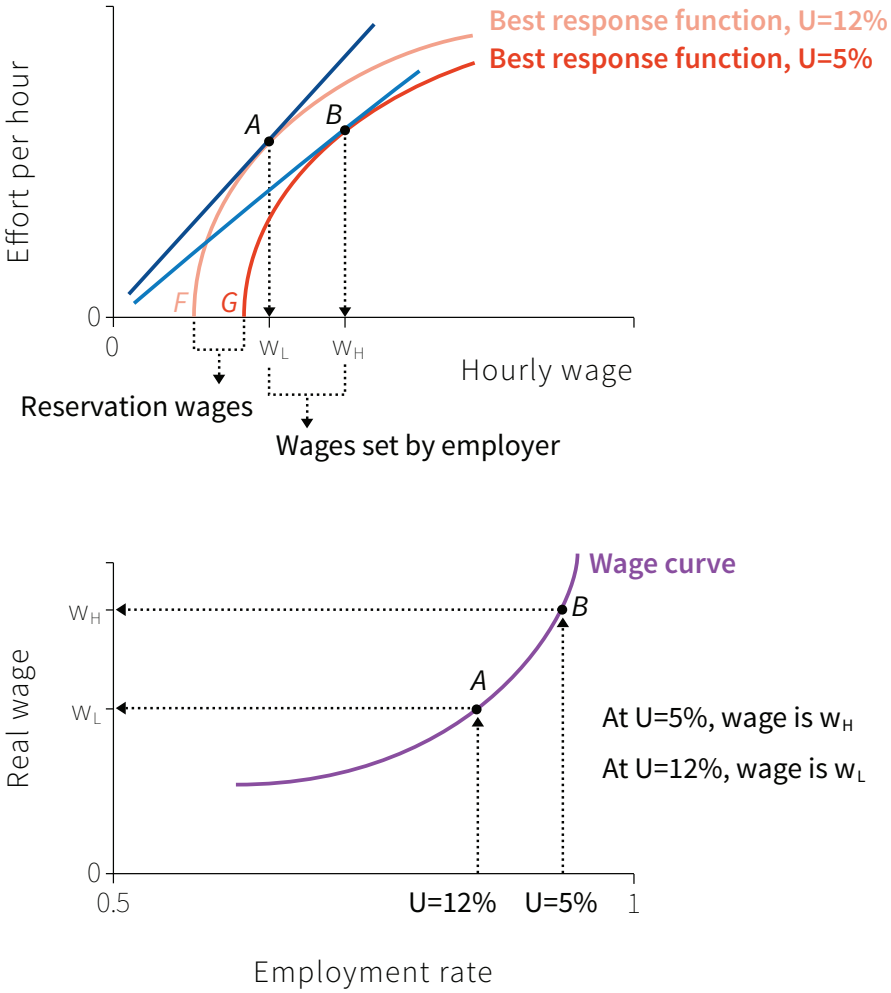


Figure 6.9 *Deriving the wage curve: varying the unemployment rate in the economy.*

In the lower panel, we plot point *A*: the dashed line from unemployment of 12% indicates that the wage, w_L is set. Remember that in the lower panel along the horizontal axis, as the employment rate increases to the right, the unemployment rate falls.

Using exactly the same reasoning, we find the profit-maximising wage set when unemployment is much lower at 5%. The reservation wage is higher and the wage set by the employer is higher as shown by point *B*. This gives the second point on the wage curve in the lower panel.

An example of how economic policy affects the wage curve is shown in Figure 6.10. Throughout this example, the unemployment rate is held constant at 12%. Rather than varying the unemployment rate, here we vary the out of work benefit to which the worker is entitled. A higher unemployment benefit increases the reservation wage and shifts the best response curve to the right: the higher reservation wage at a higher unemployment benefit level is shown by point *G*. The employer sets a higher wage (point *C*).

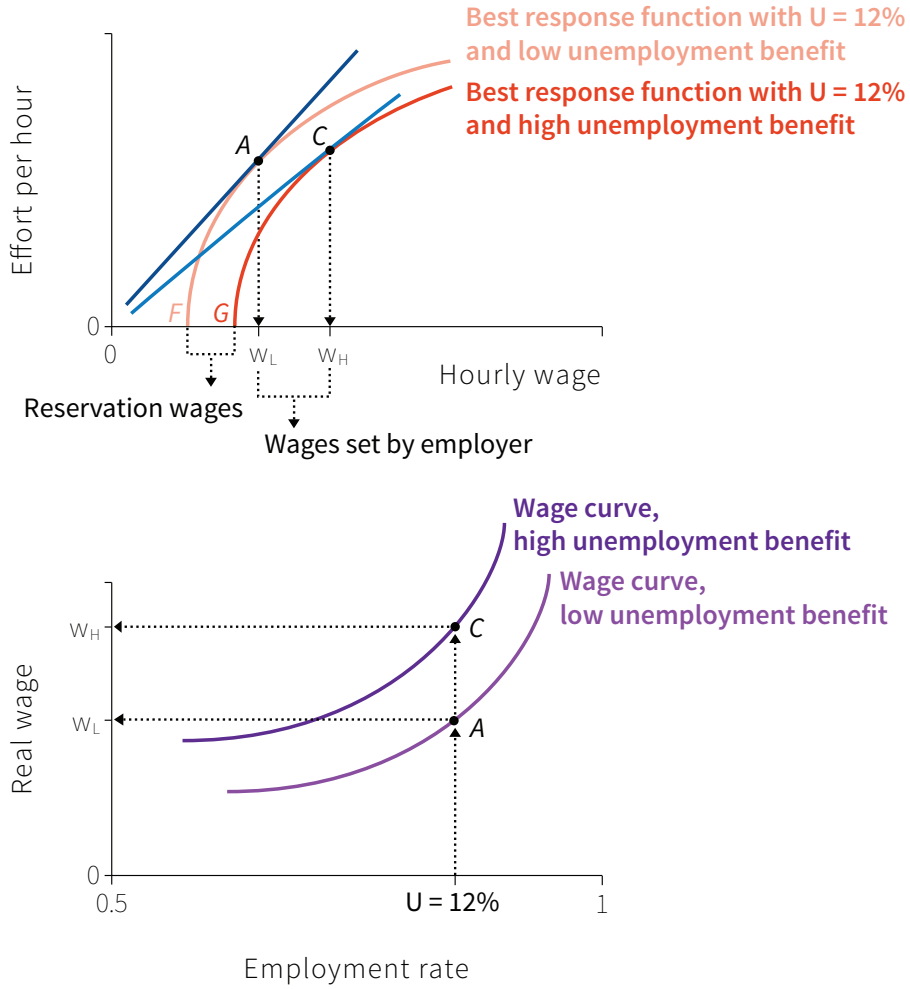


Figure 6.10 *Shifts of the wage curve: varying the unemployment benefit.*

In the lower panel, the unemployment rate is held constant at 12%. With low unemployment benefits, the wage curve goes through point A. With a high unemployment benefit, there is a new wage curve, which goes through point C.

The shape of the wage curve highlights an important limit on policies to reduce unemployment. According to our model, any policy that comes close to entirely eliminating unemployment would put employers in a position that the best they could do would be to pay wages so high they that would eliminate the employers' profits and drive the firms out of business. You have already seen in the "proof by

contradiction" that if everyone had a job then the duration of unemployment would be zero (one would immediately get another job) no matter how high a wage the employer set, so the employment rent would be zero and nobody would work.

We derived the wage curve as part of the labour discipline model designed to illuminate how employees and firm owners (and their managers) interact in setting wages and determining the level of work effort. We will use the same model later when we describe policies to alter the level of unemployment in the entire economy.

It would be valuable then to know if the facts we can get from real economies are what we would expect given the model.

Figure 6.11 is a wage curve estimated from data for the US. Note that in Figure 6.11, the horizontal axis shows the unemployment rate explicitly, falling from left to right. By using data on unemployment rates and wages in local areas, economists can estimate and plot the wage curve for an economy.

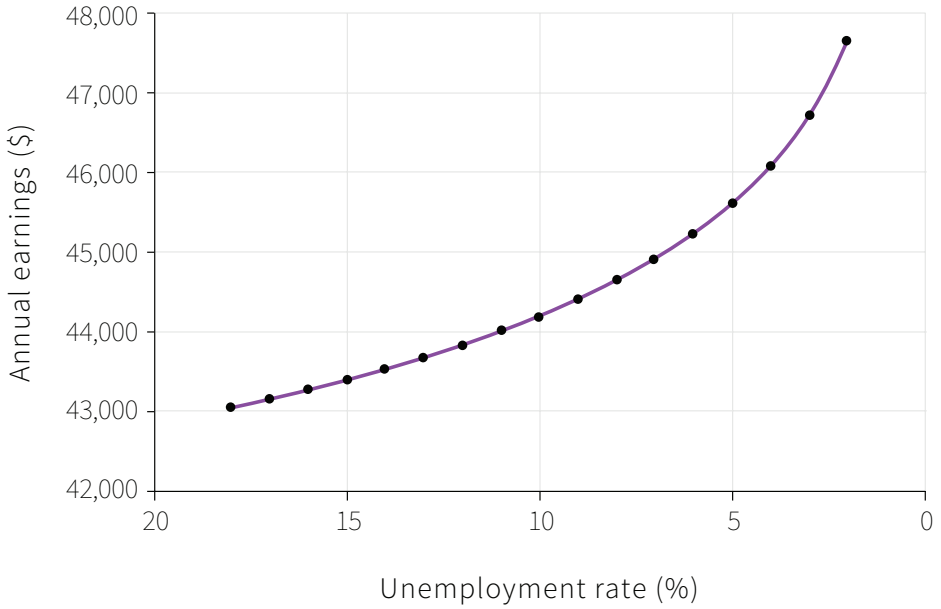


Figure 6.11 *A wage curve estimated for the United States economy (1979-2013).*

Source: *Estimated by Stephen Machin (UCL, 2015) from US Census Bureau. 2015. 'Current Population Survey', Outgoing Rotation Groups for 1979 to 2013. The sample is for males aged 26-64. Earnings are in 2013 prices. In the estimations, controls were included for the years in the labour force (experience and experience squared), gender, four levels of education, and dummy variables for states and years. The sample size is 2,346,129.*

HOW ECONOMISTS LEARN FROM FACTS

WORKERS SPEED UP WHEN THE ECONOMY SLOWS DOWN

The idea that employment rents are an incentive for employees to work harder is illustrated in a study by Edward Lazear (an economic advisor to former US President George W. Bush) and his co-authors. They investigated a single firm during the global financial crisis, to see how the managers and workers reacted to the turbulent economic conditions. The firm specialises in technology-based services, such as insurance-claims processing, computer-based test grading and technical call centres, and operates in 12 US states. The nature of the work made it easy for the management of the firm to track the productivity of workers, which is a measure of worker effort. It also allowed Lazear and his colleagues to use the firm's data from 2006-2010 to analyse the effect on worker productivity of the worst recession since the Great Depression.

When unemployment rose, workers could expect a longer spell of unemployment if they lost their job. Firms did not use their increased bargaining power to lower wages as they could have, fearing the reaction of their employees.

Lazear and his co-authors found that, in this firm, productivity increased dramatically as joblessness rose during the financial crisis. One possible explanation is that average productivity increased because management let go the least productive members of the workforce. But Lazear found that the effect was more due to workers putting in extra effort. The severity of the recession raised the workers' employment rent for any given wage, and they were therefore willing to work harder. In the model we have developed we would predict that the best response curve shifted to the left as a result of the recession. This meant that unless employers lowered wages by a large enough amount, workers would work harder. Apparently, this is what happened.

An earlier recession provided another insight that helps to explain the reluctance of the employer to reduce wages sufficiently in the crisis. Truman Bewley, an economist, was puzzled when he saw only a handful of firms in the north-east of the US cutting wages during the recession of the early 1990s. Most firms, like the one Lazear's team studied, did not cut their wages at all. Economic logic dictates that employers could have cut wages while sustaining an employment rent sufficient to motivate hard work.

Bewley interviewed more than 300 businesspeople, labour leaders, business consultants and careers advisors in the northeast of the US. He found that employers chose not to cut wages because they thought it would hurt employee morale, reducing productivity and leading to problems of hiring and retention. They thought it would ultimately cost the employer more than the money they would save in wages.

How do we show this in the model? If workers view the employer as being unfair, this could raise the disutility of work and shift the best response curve to the right. Worker effort would fall and the employer would lose. Employers thought that unless the downturn was expected to last a long time it was better to lay off some workers and keep wages the same: those who stay feel lucky to have a job, and will be willing to work a little harder, as observed by Lazear.

DISCUSS 6.7: LAZEAR'S RESULTS

Use the best response diagram to sketch the results found by Lazear and co-authors. Start by drawing three best response curves:

1. The pre-crisis period, 2006
2. The crisis year, 2007-8
3. The post-crisis5 year, 2009

Assume that the employer did not adjust wages.

6.10 LABOUR SUPPLY, LABOUR DEMAND, AND BARGAINING POWER

We began this unit with a contrast between the firm and the market, and the corresponding difference between contracts for goods and services, such as shirts and car repair on the one hand, and contracts that employers offer their employees on the other. Not surprisingly given these differences, we will see in Unit 9 that the market for labour is unlike the markets for shirts and car repair in a number of important ways.

The model of wage-setting by employers and effort provided by employees that we have studied in this unit shows how changes in the economy-wide level of unemployment affect the outcome of this process. You can visualise the number of workers lining up for jobs and the number of firms hiring some of them:

- *Supply of labour* at the going wages and working conditions is the number of people seeking work.
- *Demand for labour* is the number of jobs that employers are seeking to fill.
- The employee agrees *to follow the direction* of the employer in exchange for a wage.
- *Some degree of unemployment is necessary* for the profitable employment of labour, meaning that the supply of labour exceeds the demand for labour.

As shown in Figure 6.9, at an unemployment rate of 12%, employees are in a weak bargaining position and employers can set a relatively low wage. When demand for labour is high relative to supply with unemployment of 5%, employees are in a strong bargaining position and employers set a higher wage.

This helps explain what we saw in Unit 2 when the Black Death reduced the supply of labour in London's population. This, along with political changes at the time, drove up the real wage. The subsequent recovery of population, by increasing the supply of labour drove the wage back down again (Figure 2.13). Much later, the Industrial Revolution, along with political changes at the time altered the forces of supply and demand for labour. This increased the bargaining power of employees and drove up wages (Figure 2.14).

The model, and Figure 6.9, make it clear that the way that supply and demand affect the wage is by altering the reservation wage. An increase in the supply of labour or a decrease in the demand for labour:

- Increases the number of unemployed…
- …which increases the amount of time that a person losing a job may expect to remain unemployed…
- …which reduces the reservation wage of the employee…
- …which puts downward pressure on the wage

By similar reasoning a decrease in the supply of labour, or an increase in the demand for labour, increases the reservation wage of the employee, which puts upward pressure on the wage.

In later units we will see that changes in supply and demand alter prices of other goods and services for much the same reason that they alter wages, namely because they alter the bargaining power of buyers and sellers of shirts, car repair services and other products.

In sum:

- *Supply and demand for labour in the labour market affect the wage* (the price of an hour of working time) in much the same way as the supply and demand for cars affect the price of cars.
- *The degree of unemployment affects the bargaining power of owners and employees*, primarily by lowering the reservation wage of employees.

6.11 FAIRNESS, RECIPROCITY AND TRADE UNIONS

Our model of the firm (and the resulting model of the entire labour market) was based on some simplifications. Now we can explore the ways that both employers and employees might act differently if we loosen the restrictions of the model:

- *We now let employers introduce company policies* appealing to their employees' sense of fairness and reciprocity, so as to shift the employees' best response curve and increase profits.
- *Employees may bargain with the firm* as members of a *trade union*.

Considering the first, so far we have asked what wage the employer will offer so as to maximise profits. What other profit-maximising strategies might the employer follow? Maybe to try to move the entire best response curve upward, making possible more profitable allocations.

For example, if managers improve working conditions by providing some amenity that the worker values—air conditioning or a fair method of resolving conflicts—the job with this particular employer will be more valuable to the worker because the disutility of working there is less. As a result, the worker will work harder for any given wage as a result of the increased employment rent. The amenity could also be flexible work hours or free breakfast or free drinks after work on Fridays. The employer now needs to pay less for a given amount of effort, or equivalently can get more effort for the same wage. If this advantage more than offsets the cost of the policy, the firm will adopt it.

Figure 6.12 illustrates the effect of such a policy.

DISCUSS 6.8: USING FIGURE 6.12

1. Explain how the employer's policies altered the status quo best response function.
2. Add an employer's isoprofit line tangent to each of the three best response functions.
3. Explain why these tangencies determine the profit-maximum allocation for the employer.
4. Indicate the wage the employer will set given each of the best response functions.
5. Given each of these wages, what is the effort level of the employee?

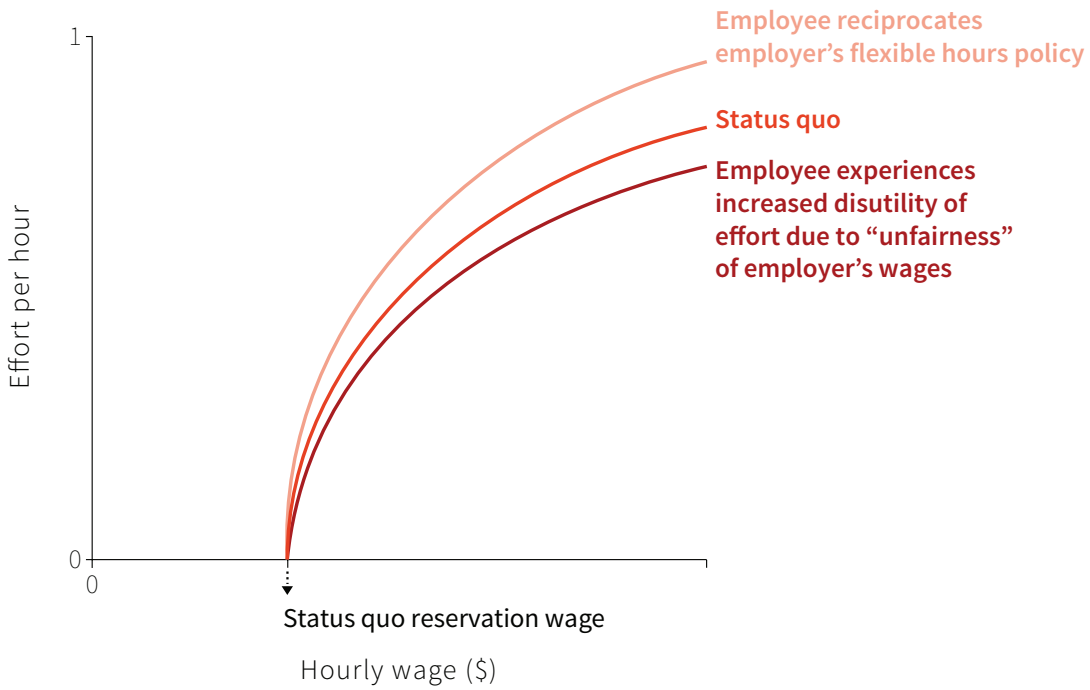


Figure 6.12 *The best response curve depends on fairness and working conditions.*

As shown in Figure 6.12, the disutility of effort may also be affected by an employee's feelings about the company, or its owners and managers. We know that from the ultimatum game in Unit 4 that people care about being treated fairly. If other similar employers have recently raised their wages, and the employee's firm has not, it is likely that workers will regard their wages as unfair. At any given wage greater than the reservation wage, the employee will as a result work less hard.

Employee concerns about fairness and feelings of reciprocity between employees and employers also help explain the effects of trade unions on work effort and wages. A trade union is an organisation that can represent the interests of a group of workers in negotiations with employers over issues such as pay, working conditions and working hours.

A union can threaten to strike or may adopt a "go slow" policy on the job in which all employees reduce their effort levels. These and other tactics that a union may follow give it some bargaining power in negotiations with employers. Thus it may be the trade union rather than the employer that sets the wage, or more likely some negotiation between the two. This means that the basic game has changed. Rather than the employer setting the wage and the employees individually responding, the sequence would now be:

1. The union sets the wage.
2. The employer informs workers that insufficient work will result in job termination.

3. Employees respond to the wage and the prospect of termination by performing some effort level.

In this case the employer no longer sets the wage that maximises profits (the point of tangency of the isoprofit line and the best response curve at point *A* in Figure 6.13a). Use the slideline to see what happens when the union sets the wage.

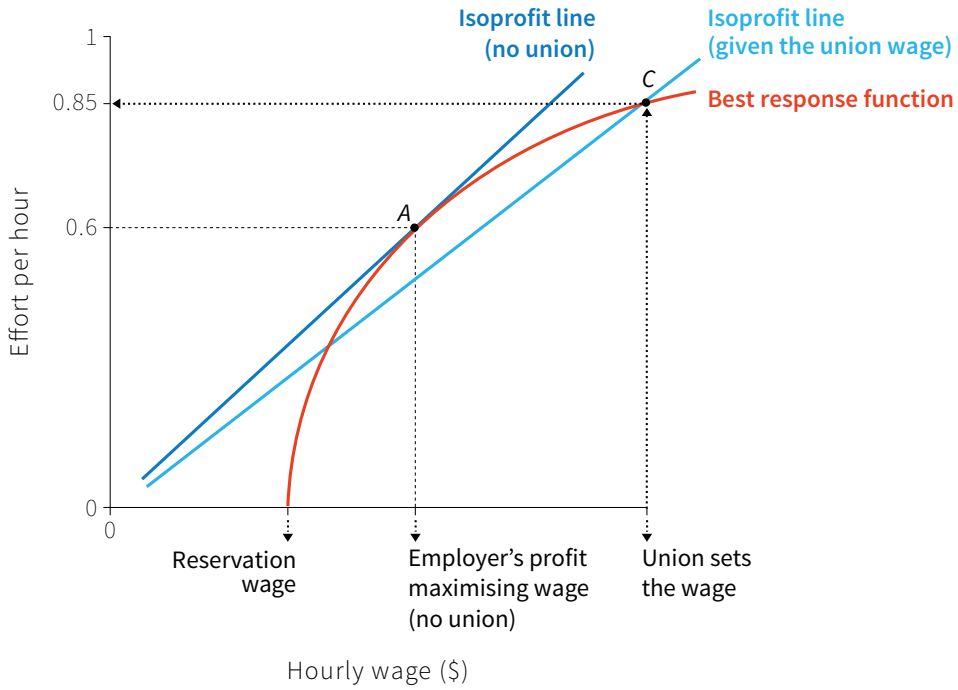


Figure 6.13a *The union sets the firm's wage.*

As shown in the figure, the wage will be higher than that preferred by the employer, but profits would be lower on the flatter isoprofit line passing through *C*. Lower profits must be the case unless the union implemented exactly the wage preferred by the owner. If nothing else changed, the owners would now receive less effort from workers for each dollar spent on wages.

But this may not be the entire story. Suppose that over the years the employer and the trade union had developed a constructive working relationship, for example solving problems that came up in ways that benefit both employees and the owners. The employees may interpret the employer's recognition of the trade union, and its willingness to compromise with them over a higher wage, as a sign of goodwill. As a result they might identify more strongly with their firm, and experience effort as less of a burden than before, shifting the best response curve in Figure 6.13b up. The result of the greater bargaining power of the workers, and their reciprocation of the company's worker-friendly policy, is shown as point *D* in the figure.

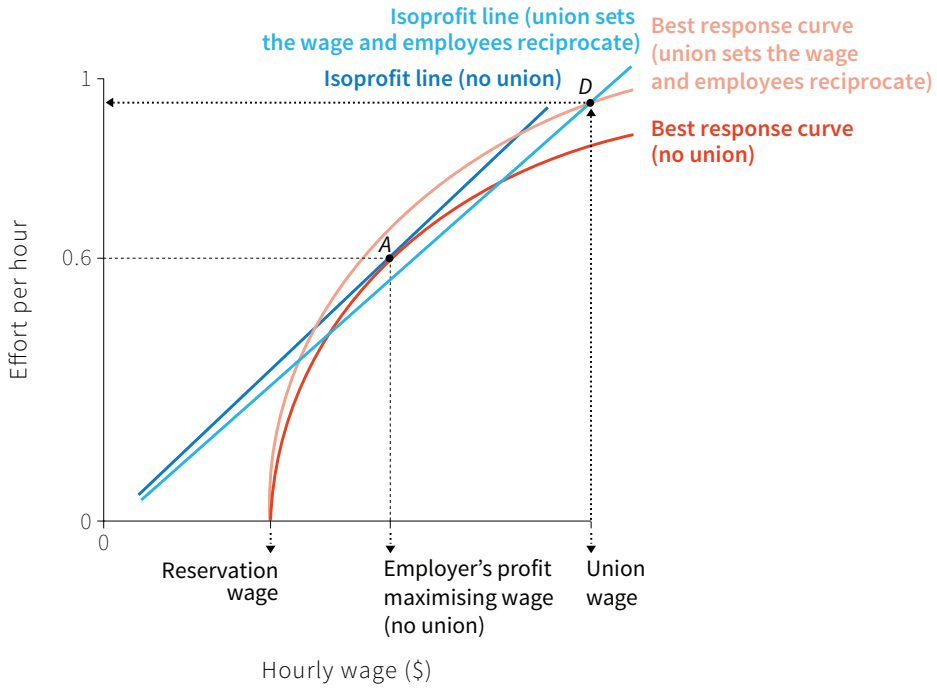


Figure 6.13b *The union sets the firm's wage and employees reciprocate.*

DISCUSS 6.9: OUTSOURCING COMES HOME

At the start of this unit we discussed the decision by many clothing companies to outsource production to Bangladesh and other low-wage economies.

1. Use the diagram with the wage on the horizontal axis and effort on the vertical axis to show the best response curve of the workers in the high-wage home country.
2. In the same diagram show the best response curve of workers in the foreign low-wage country. (Assume that there are no unions in either economy and that wages are measured in dollars in both cases.)
3. What wage does the employer set in the absence of the possibility of outsourcing?
4. What wage does the employer set if it switches production to the low-wage country (ignore the costs of moving production)?
5. How do you think the threat of outsourcing might affect the wage set by the employer in the home country and why?
6. Show this in a diagram.

6.12 ANOTHER KIND OF BUSINESS ORGANISATION

Even in capitalist economies some business organisations have an entirely different structure to the one we have been analysing: their workers are the owners of the capital goods and other assets of the company, and they select managers who run the company on a day to day basis. This form of business organisation is called a worker-owned *cooperative*.

One well-known example is the large British retailer John Lewis, founded in 1864 and held in trust for its employees since 1950. Every employee is a partner, and employee councils elect five out of seven members of the company board. The benefits for employees—pension, paid holidays, long-service sabbaticals, social activities—are generous; and the business's profits are shared out as a bonus, calculated as a percentage of each person's salary every year. The bonus normally ranges between 10% and 20% of pay even after a significant chunk of the profits are retained for future investment. John Lewis is one of the country's most profitable and consistently successful retail businesses.

Worker-owned cooperatives are hierarchically organised, like conventional firms, but the directives issued from the top of the hierarchy come from people who owe their jobs to the worker owners. Other than this, the main differences between conventional firms and worker owned cooperatives are that the cooperatives need fewer supervisors and other management personnel to ensure that the worker owners work hard and well. Fellow worker owners will not tolerate a shirking worker because the shirker is reducing the profit share of the other workers. Reduced need for the supervision of workers is among the reasons that worker-owned cooperatives produce at least as much if not more per hour than their conventional counterparts.

Inequalities in wages and salaries within the company—for example between managers and production workers—are also typically less in worker-owned cooperatives than in conventional firms. Worker-owned cooperatives also tend not to lay off workers when the economy goes into recession, offering their worker owners a kind of insurance (often they cut back on the hours of all workers rather than terminating the employment of some).

Case studies show that, in those unusual companies owned primarily by the workers themselves, work is done more intensely with less supervision. There have been many attempts to establish other types of business organisation throughout recent history, but borrowing the funds to start and sustain worker-owned companies is often difficult because, as we will see in Unit 11, banks are often reluctant to lend funds (except at high interest rates) to people who are not wealthy.

DISCUSS 6.10: A WORKER-OWNED COOPERATIVE

In Figure 6.1 we showed the actors and decision-making structure of a typical firm.

1. How do the actors and decision-making structure of John Lewis differ from that of a typical firm?
2. Redraw Figure 6.1 to show this.

GREAT ECONOMISTS

JOHN STUART MILL

John Stuart Mill (1806-1873), one of the most important philosophers and economists of the 19th century. His book *On Liberty* (1859) parallels Adam Smith's *Wealth of Nations* in advocating limits on governmental powers, and is still an influential argument in favour of individual freedom and privacy.

Mill thought that the structure of the typical firm was an affront to freedom and individual autonomy. In *The Principles of Political Economy* (1848), Mill described the relationship between firm owners and workers as an unnatural one: "To work at the bidding and for the profit of another, without any interest in the work… is not, even when wages are high, a satisfactory state to human beings of educated intelligence," he wrote.

Attributing the conventional employer-employee relationship to the poor education of the working class, he predicted that the spread of education, and the political empowerment of working people, would change this situation:

"The relation of masters and work-people will be gradually superseded by partnership… perhaps finally in all, association of labourers among themselves."
— John Stuart Mill, *The Principles of Political Economy* (1848)

DISCUSS 6.11: WAS MILL WRONG?

So far Mill's vision of a post-capitalist economy of worker-owned cooperatives has not occurred.

Why?

6.13 CONCLUSION

To understand the firm's role in the economy, we view the firm not only as an actor, but also a stage on which the actors that make up each firm—owners, managers, and employees—interact.

The three sets of actors come together in the firm because they expect to be better off participating in the firm than they would be otherwise. And they are indeed better off. We have already seen that workers earn economic rents, so they are doing better than they would without the job. The same is true of managers. Owners of the firm are of course making sufficient profits to continue to invest in this firm, instead of moving their funds elsewhere.

But wherever there are mutual gains to be had, there will be conflicts over the distribution of these gains. There are conflicts of interest between the interests of the owners (greater profits), the managers (greater salaries, first class air travel) and employees (higher wages, a safe work environment or a less punishing pace of work).

These conflicts are unavoidable because what each person receives depends on what another person does, and gets, and typically people face a situation of scarcity so that if all of the potential mutual gains have been realised, then by definition win-win outcomes are not possible.

CONCEPTS INTRODUCED IN UNIT 6

Before you move on, review these definitions:

- *Wage curve*
- *Worker's best response function*
- *Firm-specific assets*
- *Incomplete contract*
- *Employment rent*
- *Division of labour*
- *Reservation wage*
- *Trade union*

As in previous units, the institutions governing the relationships among the firm's actors, and the firm's relationship with the rest of the economy, influence whether the possible gains from exchange are fully realised and fairly distributed. We have seen that the wage rate and effort provided by an employee, for example, will depend on the bargaining power of the employees, the managers and owners, and therefore will be affected by the extent of unemployment and unemployment insurance in the economy.

In addition to social interactions among owners, managers, and employees, firms also interact with their customers. We turn now to study how firms set prices of their goods rather than the wages they pay. We will see that opportunities for mutual gain and conflicts over the division of these gains arise.

Key points in Unit 6

Markets and firms

Markets and firms are two different ways that the different products of people's labour get transferred from the producer to others (whether consumers or other producers).

Three actors in a firm

The three sets of actors benefit by participating in the firm. Their bargaining power will determine how these benefits are divided among them.

Employment contracts are incomplete

The employment contract is incomplete: it covers hours and some working conditions but not the amount of effort provided by the employee.

The employment rent

Employers set wages so that workers receive an employment rent so as to motivate workers' effort on the job and to deter employees quitting, which would impose recruitment and training costs on the employer.

Involuntary employment

There is involuntary employment because in a Nash equilibrium on the wage curve there are unemployed people who would accept a job at the prevailing wage and working conditions.

The worker's reservation wage

The supply and demand for labour affect the wage rate by affecting unemployment, which in turn alters the worker's reservation wage.

The wage curve

The wage curve for the entire economy is upward-sloping with increased employment, highlighting the fact that the employers have to pay workers a higher wage to address the triple problems of motivation, recruitment and retention when unemployment is low and hence workers reservation wage is higher.

Unemployment insurance and trade unions

Public policies such as unemployment insurance as well as trade unions and company policies can affect the wage setting process and shift the wage curve.

6.14 READ MORE

Bibliography

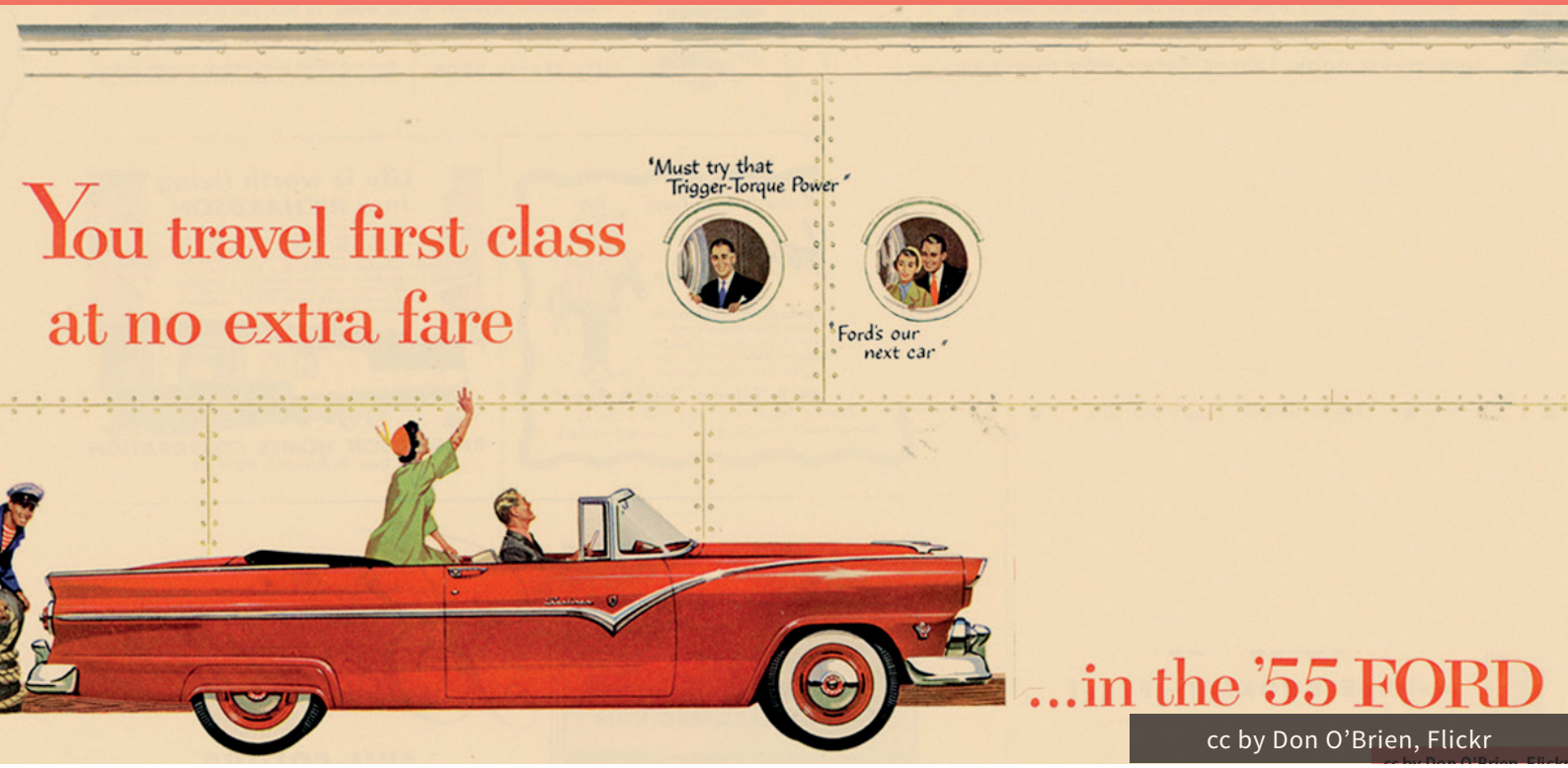
1. Bewley, Truman F. 1999. *Why Wages Don't Fall during a Recession*. Cambridge, MA: Harvard University Press.
2. Braverman, Harry, and Paul M. Sweezy. 1975. *Labor and Monopoly Capital: The Degradation of Work in the Twentieth Century*. New York, NY: Monthly Review Press.
3. Coase, Ronald H. 1937. 'The Nature of the Firm.' *Economica* 4 (16): 386–405.
4. Coase, Ronald H. 1992. 'The Institutional Structure of Production.' *American Economic Review* 82 (4): 713–19.
5. Couch, Kenneth A., and Dana W. Placzek. 2010. 'Earnings Losses of Displaced Workers Revisited.' *American Economic Review* 100 (1): 572–89.
6. Ehrenreich, Barbara. (2001) 2011. *Nickel and Dimed: On (Not) Getting By in America*. New York, NY: St. Martin's Press.
7. Friedman, Milton. 1970. 'The Social Responsibility of Business Is to Increase Its Profits.' *New York Times*, September.
8. Hansmann, Henry. 2000. *The Ownership of Enterprise*. Cambridge, MA: Belknap Press.
9. Helper, Susan, Morris Kleiner, and Yingchun Wang. 2010. 'Analyzing Compensation Methods in Manufacturing: Piece Rates, Time Rates, or Gain-Sharing?' *NBER Working Papers* No 16540, National Bureau of Economic Research, Inc.
10. Jacobson, Louis, Robert J. Lalonde, and Daniel G. Sullivan. 1993. 'Earnings Losses of Displaced Workers.' *The American Economic Review* 83 (4): 685–709.
11. Kletzer, Lori G. 1998. 'Job Displacement.' *Journal of Economic Perspectives* 12 (1): 115–36.
12. Kroszner, Randall S., and Louis Putterman, eds. 2009. *The Economic Nature of the Firm: A Reader*. Cambridge: Cambridge University Press.
13. Krueger, Alan B., and Alexandre Mas. 2004. 'Strikes, Scabs, and Tread Separations: Labor Strife and the Production of Defective Bridgestone/Firestone Tires.' *Journal of Political Economy* 112 (2): 253–89.
14. Lazear, Edward P., Kathryn L. Shaw, and Christopher Stanton. 2013. 'Making Do with Less: Working Harder During Recessions.' *NBER Working Papers* No 19328, National Bureau of Economic Research Inc.
15. Marx, Karl. (1867) 1906. *Capital: A Critique of Political Economy*. New York, NY: Random House.
16. Marx, Karl. (1848) 2010. *The Communist Manifesto*. Edited by Friedrich Engels. London: Arcturus Publishing.

17. Micklethwait, John, and Adrian Wooldridge. 2003. *The Company: A Short History of a Revolutionary Idea*. New York, NY: Modern Library.
18. Mill, John Stuart. (1848) 1994. *Principles of Political Economy*. New York: Oxford University Press.
19. Mill, John Stuart. (1859) 2002. *On Liberty*. Mineola, NY: Dover Publications.
20. O'Reilly, Tim, and Eric S. Raymond. 2001. *The Cathedral & the Bazaar: Musings on Linux and Open Source by an Accidental Revolutionary*. Sebastopol, CA: O'Reilly.
21. Pencavel, John. 2002. *Worker Participation: Lessons from the Worker Co-Ops of the Pacific Northwest*. New York, NY: Russell Sage Foundation Publications.
22. Seabright, Paul. 2010. *The Company of Strangers: A Natural History of Economic Life*. Princeton, NJ: Princeton University Press.
23. Simon, Herbert A. 1951. 'A Formal Theory of the Employment Relationship.' *Econometrica* 19 (3).
24. Simon, Herbert A. 1991. 'Organizations and Markets.' *Journal of Economic Perspectives* 5 (2): 25–44.
25. US Census Bureau. 2015. 'Current Population Survey.'
26. Williamson, Oliver E. 1985. *The Economic Institutions of Capitalism*. New York, NY: Collier Macmillan.

Beta September 2015 version

7

THE FIRM AND ITS CUSTOMERS



HOW A PROFIT-MAXIMISING FIRM PRODUCING A DIFFERENTIATED PRODUCT INTERACTS WITH ITS CUSTOMERS

- Differentiated products, the product demand curve and the firm's marginal cost
- Technological and cost advantages of large-scale production favour large firms
- How a firm without close competitors chooses the price and quantity that maximises its profits and how it increases its profits through product selection and advertising
- How the gains from trade are divided between consumers and owners of the firm
- The responsiveness of consumers to a price change is measured by the elasticity of demand, which affects the firm's price and profit margin
- How economic policymakers use elasticity of demand to design tax and competition policy

See www.core-econ.org for the full interactive version of *The Economy* by The CORE Project. Guide yourself through key concepts with clickable figures, test your understanding with multiple choice questions, look up key terms in the glossary, read full mathematical derivations in the Leibniz supplements, watch economists explain their work in Economists in Action – and much more.

Institute for New Economic Thinking
UCL SciencesPo Azim Premji University
FRIENDS PROVIDENT Foundation

Ernst F. Schumacher's *Small is Beautiful*, published in 1973, advocated small-scale production by individuals and groups in an economic system designed to emphasise happiness rather than profits. In the year the book was published the firms Intel and FedEx each employed only a few thousand people in the US; 40 years later Intel employed around 108,000, and FedEx more than 300,000 people. In 1973 Walmart employed 4,500 people; in 2014 it employed 2.2 million.

Most firms in the US are much smaller than this, but in all of the rich economies most people work for large firms. In the US, half of private sector employees work in firms with at least 1,000 employees. The main reason is that owners of firms make more money if they can expand to a larger size, and people with money to invest get higher returns from owning stock in large firms. Employees in large firms are also paid more. Figure 7.1 shows the growth of some highly successful US firms.

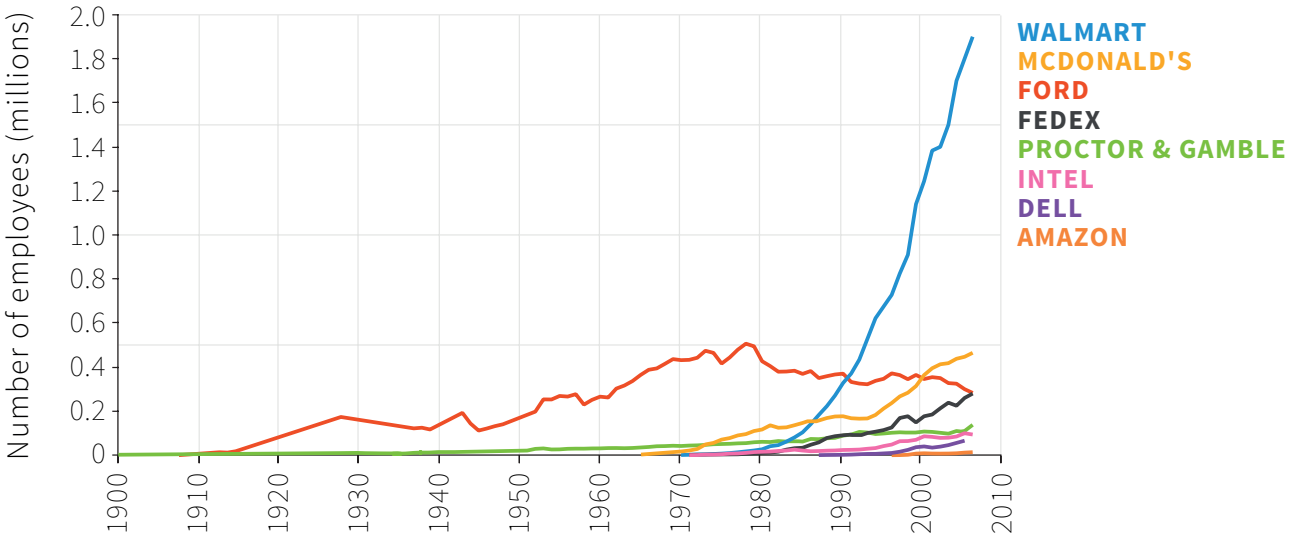


Figure 7.1 *Firm size in the United States: number of employees (1900-2006).*

Source: *Luttmer, Erzo G. J. 2011. 'On the Mechanics of Firm Growth.' The Review of Economic Studies 78 (3): 1042–68.*

"Pile it high and sell it cheap" was the motto of Jack Cohen. He started as a street trader, and founded a small supermarket called Tesco in 1919. Today, £1 in every £7 spent in a shop in the UK is spent in a Tesco store, and the company expanded worldwide in the 1990s; in 2014 Tesco had higher profits than any other retailer in the world except Walmart. Keeping the price low as Cohen recommended is one possible strategy for a firm seeking to maximise its profits: the profit on each item is small, but the low price may attract so many customers that total profit is high.

Other firms may adopt quite different strategies; Apple sets high prices for iPhones and iPads, increasing its profits by charging a price premium, rather than lowering prices to reach more customers. For example, between April 2010 and March 2012, profit per unit on Apple iPhones was between 49% and 58% of the price. During the same period, Tesco's operating profit per unit was between 6.0% and 6.5%.

Firms' success depends on more than getting the price right. Their choice of products also matters, as does their ability to attract customers, and produce at lower cost and higher quality than their competitors. They need to be able to recruit and retain employees who can make all these things happen.

Figure 7.2 illustrates key decisions that a firm makes. In this unit we will focus particularly on how a firm chooses the price of a product, and the quantity to produce. This will depend on the demand it faces—that is, the willingness of potential consumers to pay for its product, and the costs of production.

The demand for the product will depend on its price; and the costs of production may depend on how many units are produced. But a firm can actively influence both consumer demand and costs in more ways than by price and quantity produced. As we saw in Unit 2, innovation may lead to new and attractive products, or to lower production costs. If the firm can innovate successfully it can earn rents; at least in the short term until others catch up. Further innovation may be needed if it is to stay ahead. Advertising can increase demand. And as we saw in Unit 6, the firm sets the wage, which is an important component of its cost; and as we will see in later units, the firm also spends to influence taxes and environmental regulation.

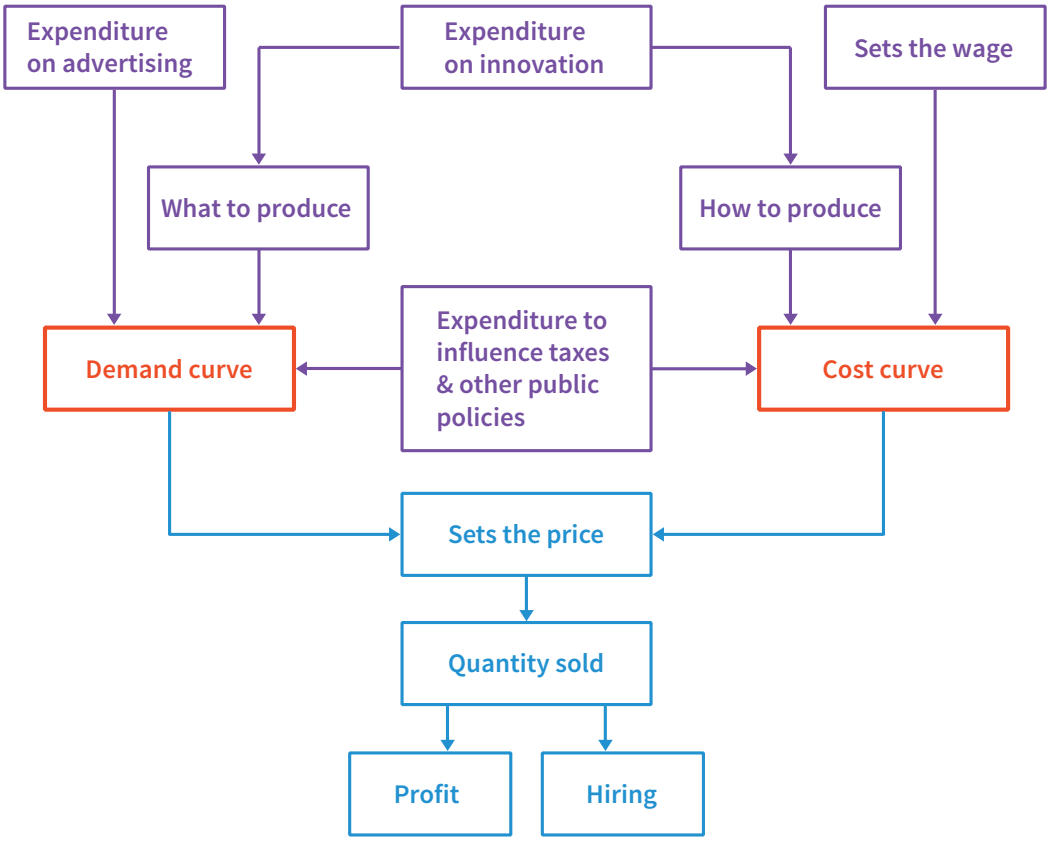


Figure 7.2 *The firm's decisions.*

7.1 BREAKFAST CEREAL: CHOOSING A PRICE

To decide what price to charge a firm needs information about demand: how much potential consumers are willing to pay for its product. Figure 7.3 shows the demand curve for Apple-Cinnamon Cheerios, a ready-to-eat breakfast cereal introduced by the company General Mills in 1989. In 1996, Jerry Hausman, an economist, used data on weekly sales of family breakfast cereals in US cities to estimate how the quantity of cereal customers would wish to buy, per week in a typical city, would vary with the price per pound weight (there are 2.2 pounds in 1kg). You can see from Figure 7.3 that, if the price were $3 for example, customers would demand 25,000 pounds of Apple-Cinnamon Cheerios. The lower the price, the more customers wish to buy, whether the product is cereal or space flight.

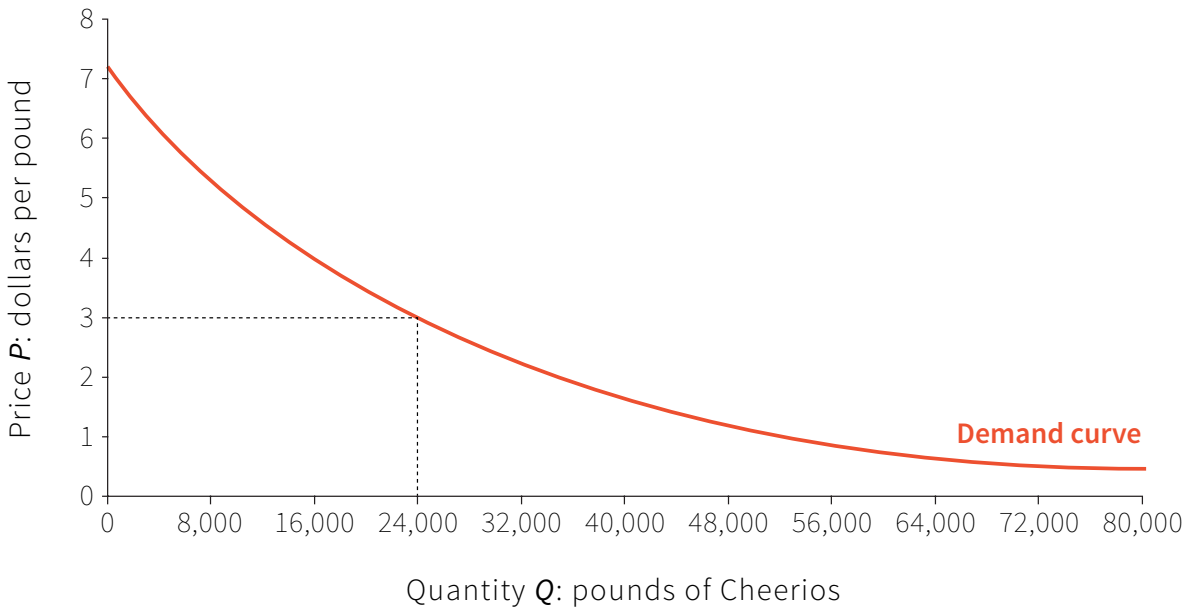


Figure 7.3 *Estimated demand for Apple-Cinnamon Cheerios.*

Source: *Adapted from Figure 5.2 in Hausman, Jerry A. 1996. 'Valuation of New Goods under Perfect and Imperfect Competition.' In The Economics of New Goods, by Robert J. Gordon and Timothy F. Bresnahan, 207–48. Chicago, IL: University of Chicago Press.*

If you were the manager at General Mills, how would you choose the price for Apple-Cinnamon Cheerios in this city, and how many pounds of cereal would you produce? You need to consider how the decision will affect your profits—the difference between sales revenue and production costs. Suppose that the unit cost—the cost of producing each pound—of Apple-Cinnamon Cheerios is $2. To maximise your profit you should produce the quantity you expect to sell, and no more. Then revenue, costs, and profit are given by:

$$\begin{aligned} \textit{total costs} &= \textit{unit cost} \times \textit{quanitity} \\ &= 2 \times Q \\ \textit{total revenue} &= \textit{price} \times \textit{quantity} \\ &= P \times Q \\ \textit{profit} &= \textit{total revenue} - \textit{total costs} \\ &= P \times Q - 2 \times Q \end{aligned}$$

So we have a formula for profit:

$$\textit{profit} = (P - 2) \times Q$$

Using this formula, you could calculate the profit for any choice of price and quantity, and draw the isoprofit curves, as in Figure 7.4. Just as indifference curves join points in a diagram giving the same level of utility, isoprofit curves join points that give the same level of profit. We can think of the isoprofit curves as your indifference curves: you are indifferent between combinations of price and quantity that give you the same profit.

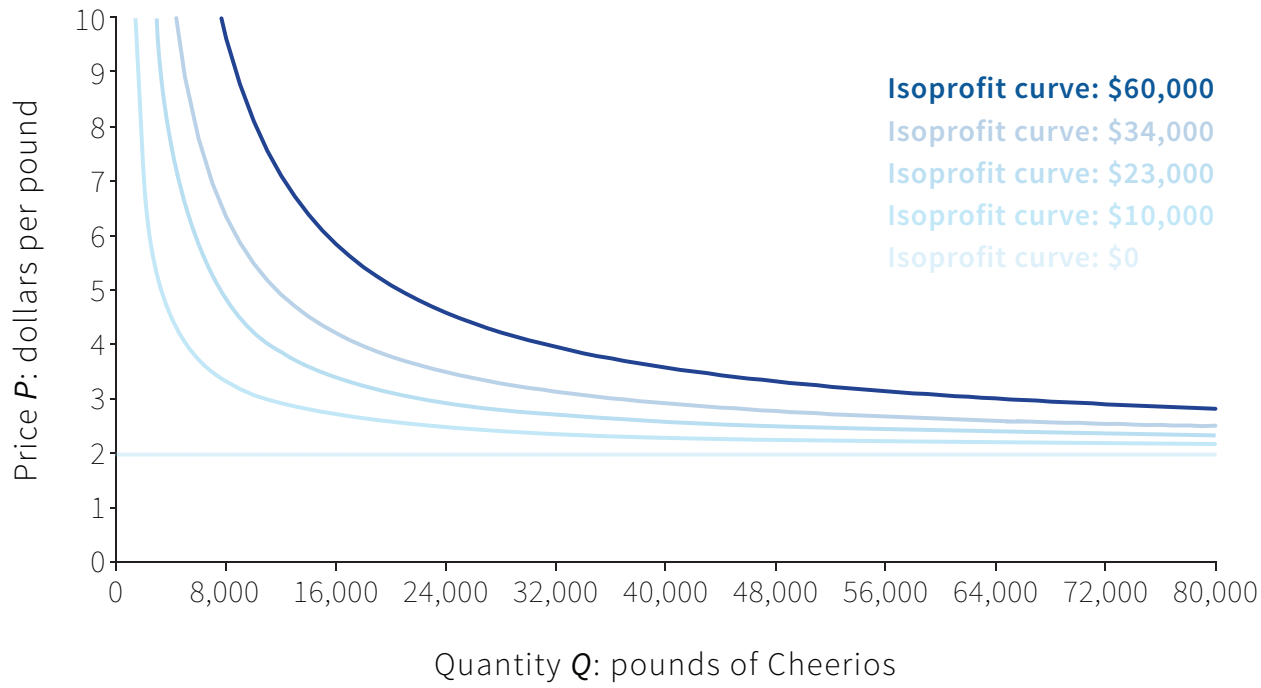


Figure 7.4 *Isoprofit curves for the production of Apple-Cinnamon Cheerios.*

Source: *Isoprofit data is illustrative only, and does not reflect the real-world profitability of the product.*

The graph shows a number of isoprofit curves for Cheerios. You could make $60,000 profit by selling 60,000 pounds at a price of $3, or 20,000 pounds at $5, or 10,000 pounds at $8, or in many other ways. The curve furthest from the origin shows all the possible ways of making $60,000 profit. The $34,000 isoprofit curve shows all the combinations of *P* and *Q* for which profit is equal to $34,000. The isoprofit curves nearer to the origin correspond to lower levels of profit. The cost of each pound of Cheerios is $2, so *profit* = *(P – 2) x Q*. This means the isoprofit curves slope downward. To make a profit of $10,000 P would have to be very high if Q was less than 8,000. But if Q = 80,000 you could make this profit with a low P. The horizontal line shows the choices of price and quantity where profit is zero: if you set a price of $2, you would be selling each pound of cereal for exactly what it cost.

To achieve a high profit, you would like both price and quantity to be as high as possible, but you are constrained by the demand curve: if you choose a high price you will only be able to sell a small quantity; and if you want to sell a large quantity, you must choose a low price. Figure 7.5a shows the isoprofit curves and demand curve together. You face a similar problem to Alexei, the student in Unit 3, who wanted to choose the point in his feasible set where his utility was maximised. You want to choose a feasible combination of price and quantity that will maximise your profit.

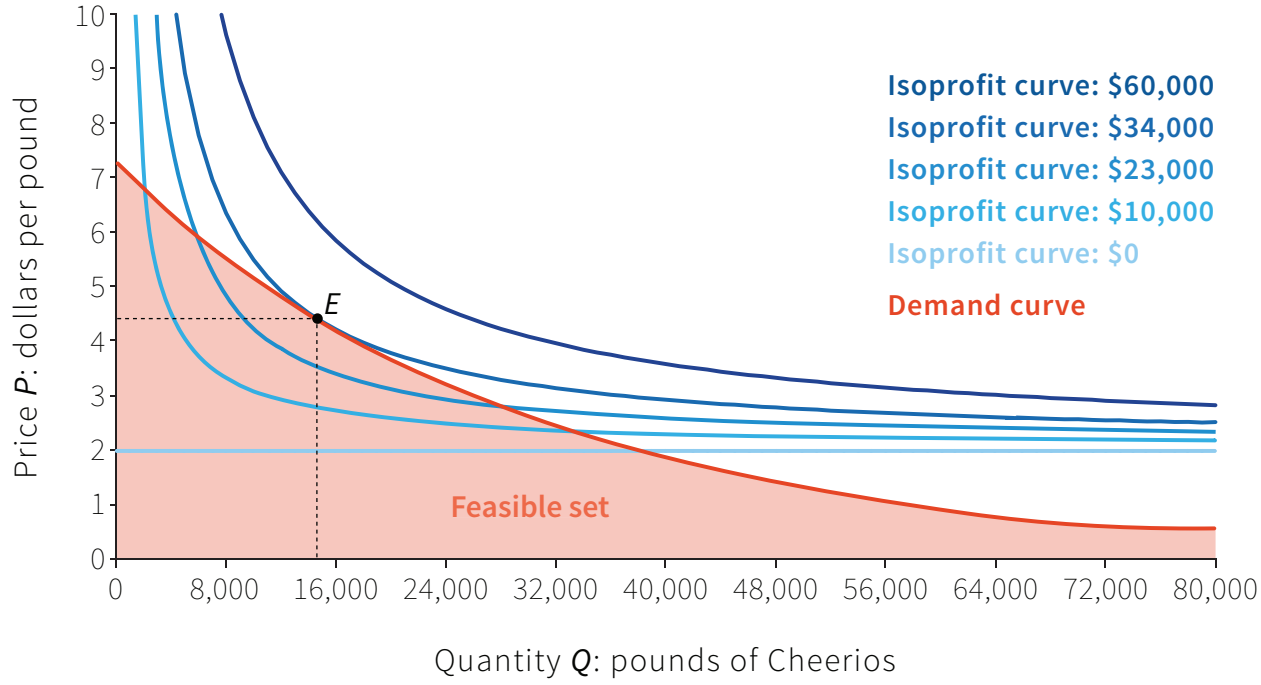


Figure 7.5a *The profit-maximising choice of price and quantity for Apple-Cinnamon Cheerios.*

Source: *Isoprofit data is illustrative only, and does not reflect the real-world profitability of the product. Demand curve data from Hausman, Jerry A. 1996. 'Valuation of New Goods under Perfect and Imperfect Competition.' In The Economics of New Goods, by Robert J. Gordon and Timothy F. Bresnahan, 207–48. Chicago, IL: University of Chicago Press.*

Your best strategy is to choose point E in Figure 7.5a: you should produce 14,000 pounds of cereal, and sell it at a price of $4.40 per pound, making $34,000 profit. Just as in the case of Alexei in Unit 3, your optimal combination of price and quantity involves balancing two trade-offs. As manager, what you care about (we have assumed) is profit, rather than any particular combination of price and quantity. The slope of the isoprofit curve at any point represents the trade-off you are *willing* to make between P and Q—your MRS. You would be willing to substitute a high price for a lower quantity if you obtained the same profit. The slope of the demand curve is the trade-off you are *constrained* to make—your MRT, or the rate at which the demand curve allows you to "transform" quantity into price. You cannot raise the price without lowering the quantity, because fewer consumers will buy. At the choice of P and Q that maximises your profit the two trade-offs balance.

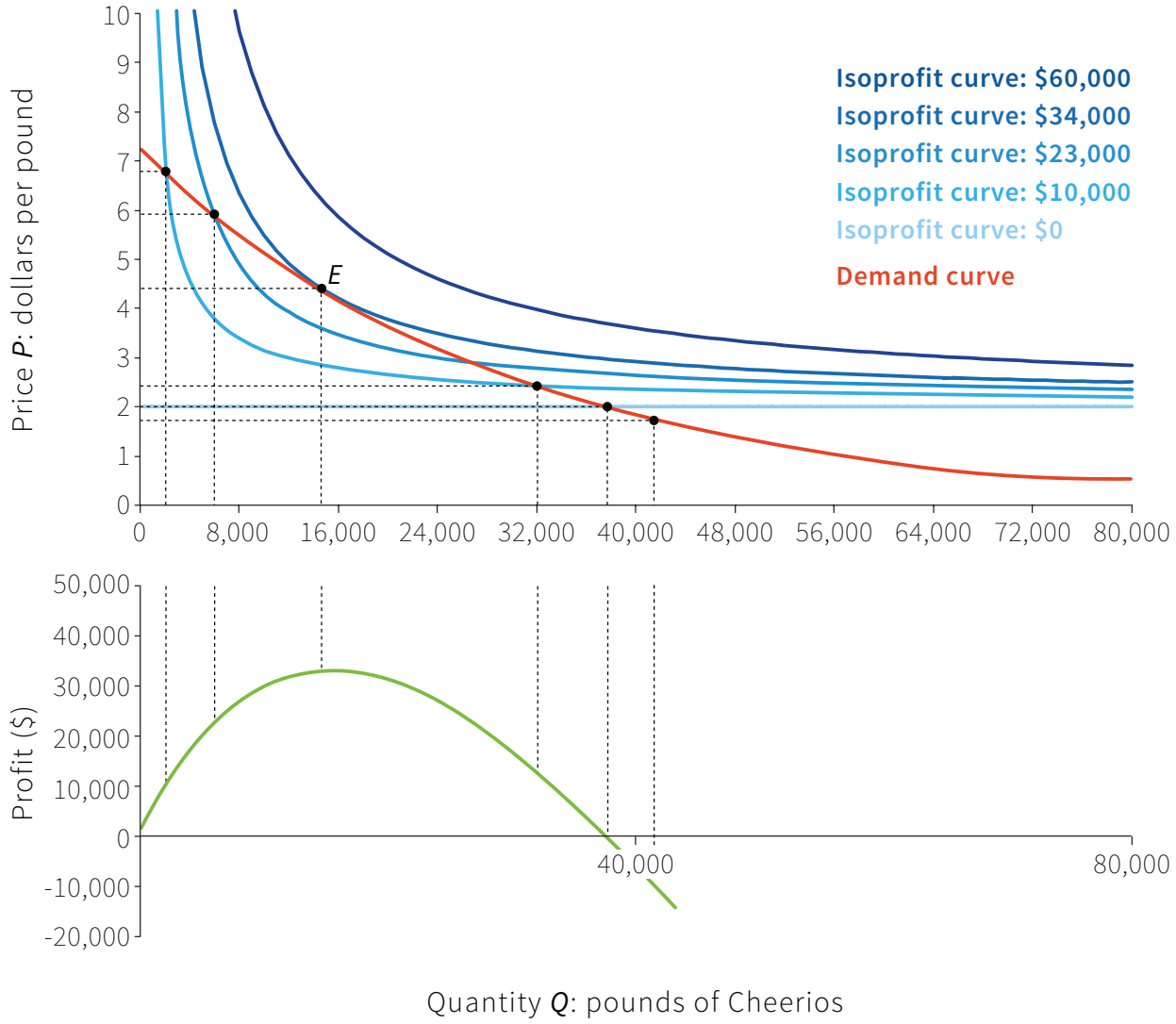


Figure 7.5b *The profit-maximising choice of price and quantity for Apple-Cinnamon Cheerios.*

Source: *Isoprofit data is illustrative only, and does not reflect the real-world profitability of the product. Demand curve data from Hausman, Jerry A. 1996. 'Valuation of New Goods under Perfect and Imperfect Competition.' In The Economics of New Goods, by Robert J. Gordon and Timothy F. Bresnahan, 207–48. Chicago, IL: University of Chicago Press.*

The manager at General Mills probably didn't think about the decision in this way. Perhaps the price was chosen more by trial and error, informed by past experience and market research. But we expect that a firm will find its way, somehow, to a profit-maximising price and quantity. The purpose of our economic analysis is not to model the manager's thought process, but to understand the outcome, and its relationship to the firm's cost and consumer demand.

Even from the point of view of the economist, there are other ways to think about profit maximisation. The lower panel of Figure 7.5b shows how much profit would be made at each point on the demand curve.

The graph in the lower panel is the profit function: it shows the profit you would achieve if you chose to produce a quantity, Q, and from the demand function, set the highest price that would enable you to sell that quantity. And it tells us, again, that you would achieve the maximum profit of $34,000 with Q = 14,000 pounds of cereal.

7.2 ECONOMIES OF SCALE AND THE COST ADVANTAGES OF LARGE-SCALE PRODUCTION

Why have firms like Walmart, Intel and Fedex grown so large? An important reason that a large firm may be more profitable than a small firm is that the large firm produces its output at lower cost per unit. This may be possible for two reasons:

- *Cost advantages*: Larger firms may be able to purchase their inputs at lower cost because they have bargaining power when they negotiate with suppliers.
- *Technological advantages*: Large-scale production often uses fewer inputs per unit of output.

Economists use the term *economies of scale* or *increasing returns* to describe the technological advantages of large-scale production. For example, if doubling the amount of every input that the firm uses triples the firm's output, then the firm exhibits increasing returns.

Economies of scale may result from specialisation among members of the firm, allowing employees to do the task at which they are best, and minimising training time by requiring a more limited skill set of employees. Economies of scale may also occur for purely engineering reasons: transporting more of a liquid requires a larger pipe, but doubling the capacity of the pipe increases the diameter of the pipe (and the material necessary to construct it) by much less than a factor of two. For proof, check *The size and cost of a pipe* in this unit's Einstein section.

ECONOMIES AND DISECONOMIES OF SCALE

If we increase all inputs by a given proportion, and it:

- Increases output more than proportionally, the technology is said to exhibit *increasing returns to scale in production* or *economies of scale*
- Increases output less than proportionally, the technology exhibits *decreasing returns to scale in production* or *diseconomies of scale*
- Increases output proportionally, the technology exhibits *constant returns to scale in production*

But there are also built-in diseconomies of scale. Think of the firm's owners, managers, work supervisors and production workers. Suppose that each supervisor directs 10 production workers, while each manager directs 10 supervisors, and so on. If the firm employs 10 production workers, then the owner can do the management and supervision. If it employs 100 production workers it then needs to add a layer of 10 supervisors and, if it grows to 1,000 production workers, it will need to recruit yet another layer of management to supervise the first layer of supervisors. If this is the case, then supervising production workers requires more than a proportional increase in the input of supervisors. The only way the firm could increase all inputs proportionally would be to reduce the intensity of supervision, with associated losses in productivity. Let's call this diseconomy of scale the *Dilbert law of firm hierarchy*, (after this comic strip). See the Einstein section for how to calculate the size of the diseconomy of scale that our Dilbert law implies.

Cost per unit may also fall as the firm produces more output, even if there are constant or even decreasing returns to scale. This will occur if there is a fixed cost that is required for the firm to produce even a single unit, which then does not increase for additional units. An example would be the cost of *research & development* (R&D) and product design, acquiring a licence to engage in production, or obtaining a patent for a particular technique. Marketing expenses such as advertising are another fixed cost. A 30-second advertisement during the television coverage of the US Super Bowl football game in 2014 cost $4m, a cost that would be justifiable only if the product would sell a large number of units as a result.

The cost of a firm's attempt to gain favourable treatment by government bodies through lobbying, contributions to election campaigns and public relations expenditures are also a kind of fixed cost. These expenses are incurred more or less independently of the level of the firm's output. We return to the question of fixed costs in Unit 8.

Also, as we have seen, large firms are able to purchase their inputs on more favourable terms than smaller firms.

Large size may also benefit a firm in selling its product, not just in producing it. This occurs when people are more likely to buy a product or service if it has a lot of users already. For example, a software application is more useful when everybody is using a compatible version. These demand-side economies of scale are called *network effects*, and there are many examples in technology-related markets.

Because producing large amounts creates economies of scale, reduces costs and increases demand, large-scale production is a powerful influence on firm size. Often production by a small group of people is simply too costly to compete with larger firms.

But while small firms typically either grow or die, there are limits to firm growth. We have already seen in Unit 6 that firms face a "make it or buy it" choice when it comes to components. The relative cost of the two options determines the choice. Firm growth is limited, in part, because sometimes it is cheaper to purchase part of the product than to manufacture it. Apple would be gigantic had it decided that Apple employees would produce the touch screens, chipsets and other components that make up the iPhone and iPad, rather than purchasing these parts from Toshiba, Samsung and other suppliers. Apple's outsourcing strategy limits the firm's size, and increases the size of Toshiba, Samsung and other firms that produce Apple's components.

In the next section, we will look at how to model the way that a firm's costs depend on its scale of production.

DISCUSS 7.1: FALL OF FORD

Compare the recent problems at Ford, the American car manufacturer, to that of the other firms in Figure 7.1.

What might explain the difference in the trends?

7.3 PRODUCTION: THE COST FUNCTION FOR *BEAUTIFUL CARS*

To set the price and the level of production for Apple-Cinnamon Cheerios the manager needed to know the demand function, and the production costs. Since we assumed that the cost of producing every pound of Cheerios was the same, the

scale of production was determined by the demand for the good. In this section and the next, we will look at a different example, in which costs vary with the level of production.

Consider a firm that manufactures cars. Compared with Ford, which produces around 6.3 million vehicles a year, this firm produces specialty cars and will turn out to be rather small—so we will call it *Beautiful Cars*.

Think about the costs of producing and selling cars. The firm needs premises—a factory—equipped with machines for casting, forging, assembling and welding car bodies. It may rent them from another firm, or raise financial capital in order to invest in premises and equipment for itself. Then it must purchase the raw materials and components, and pay production workers to operate the equipment. Other workers will be needed to manage the production process and market and sell the finished cars.

The firm's owners—the shareholders—would not be willing to invest in the firm if they could make better use of their money by investing and earning profits elsewhere. What they could receive if they invested elsewhere, per dollar of their investment, is another example of opportunity cost that we encountered in Unit 3, this time the *opportunity cost of capital*. One component of the cost of producing cars is the amount that has to be paid out to shareholders to cover the opportunity cost of capital—that is, to induce them to continue to invest in the assets the firm needs to produce cars.

The more *Beautiful Cars* are produced, the higher the total costs will be. The upper panel of Figure 7.6 shows how total costs might depend on the quantity of cars, Q, produced per day. This is the firm's cost function, $C(Q)$. From the cost function, we have worked out the average cost of a car, and how it changes with Q; the average cost curve (AC) is plotted in the lower panel.

We can see in Figure 7.6 that *Beautiful Cars* has decreasing average costs at low levels of production. The AC curve slopes downward. At higher levels of production, average cost increases so the AC curve slopes upward. This might happen because the firm has to increase the number of shifts per day on the assembly line. Perhaps it has to pay overtime rates, and equipment breaks down more frequently when the production line is working for longer.

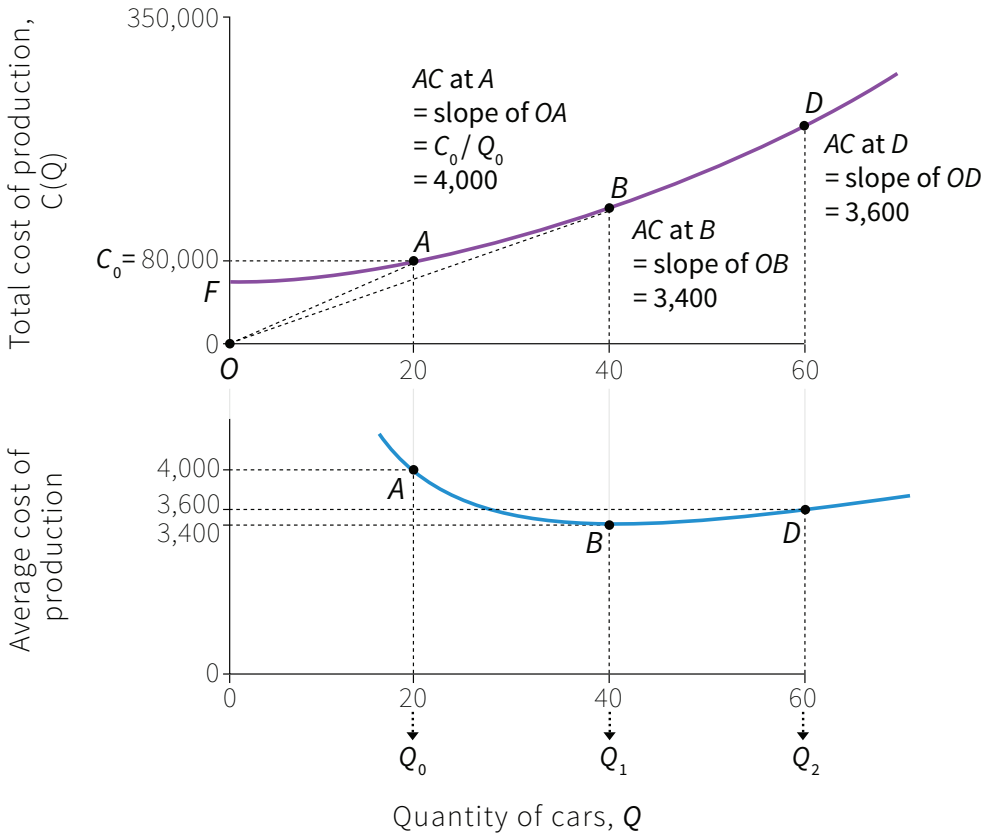


Figure 7.6 *The firm's cost function and average costs.*

The top panel shows the cost function, *C(Q)*. It shows the total cost for each level of output, *Q*. Some costs do not vary with the number of cars: for example, once the firm has decided the size of factory and invested in equipment, those costs will be the same irrespective of output. These are called *fixed costs*. So when $Q = 0$, the only costs are the fixed costs, *F*. As *Q* increases, total costs rise: the firm needs to employ more production workers. At point *A*, 20 cars are produced (we call this Q_0) costing \$80,000 (we call this C_0). If the firm produces 20 cars per day, the average cost of a car is C_0 divided by Q_0, which is shown by the slope of the line from the origin to *A*: average cost is *\$80,000/20 = \$4,000*. We have plotted the average cost at point A on the lower panel. As output rises above *A* total costs rise, but the average cost—the cost per car produced – falls. At point *B*, with output of 40 cars, the total cost is \$136,000 so the average cost has fallen to \$3,400. The fixed costs are shared between more cars. Average cost is lowest at point *B*. When production increases beyond point *B*, the line showing the average cost gets steeper again. The average cost rises. At point *D*, 60 cars are produced and the average cost has risen to \$3,600. If we calculate the average cost at every value of *Q*, we can draw the average cost *(AC)* curve in the lower panel.

Figure 7.7 shows how to find the *marginal cost* of a car: that is, the cost of producing one more car. In Unit 3 we saw that for a production function, the marginal product was the additional output produced when the input was increased by one unit, corresponding to the slope of the production function. Figure 7.7 demonstrates that the marginal cost *(MC)* corresponds to the slope of the cost function.

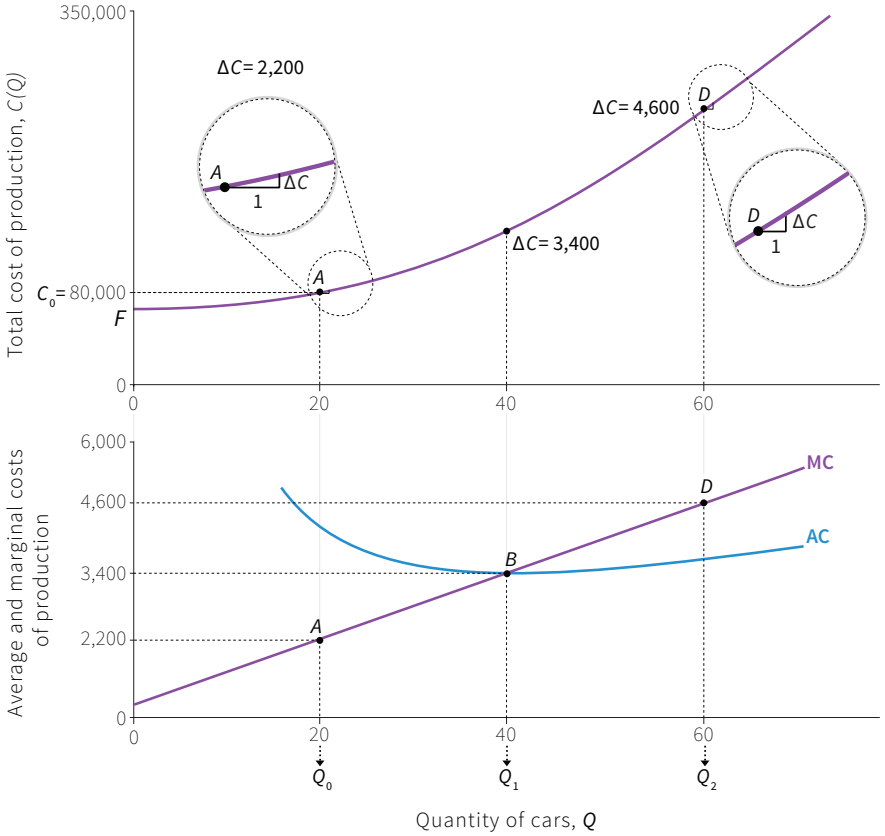


Figure 7.7 *The marginal cost of a car.*

The upper panel shows the cost function. The lower panel shows the average cost curve. We will plot the marginal costs in the lower panel too. Suppose the firm is producing 20 cars at point *A*. The total cost is $80,000. The marginal cost is the cost of increasing output from 20 to 21. This would increase total costs by an amount that we call Δ*C*, equal to $2,200. The triangle drawn at *A* shows that the marginal cost is equal to the slope of the cost function at that point. We have plotted the marginal cost at point *A* in the lower panel. At point *D*, where *Q = 60*, the cost function is much steeper. The marginal cost of producing an extra car is higher: Δ*C = $4,600*. At point *B*, the curve is steeper than at *A*, but flatter than at *D: MC = $3,400*. Look at the shape of the whole cost function. When *Q = 0* it is quite flat, so marginal cost is low. As *Q* increases, the cost function gets steeper, and marginal cost gradually rises. If we calculate marginal cost at every point on the cost function, we can draw the marginal cost curve.

By calculating the marginal cost at every value of Q, we have drawn the whole of the marginal cost curve in the lower panel of Figure 7.7. Since *MC* is the slope of the cost function and the cost curve gets steeper as Q increases, the graph of marginal cost is an upward-sloping line. In other words, *Beautiful Cars* has increasing marginal costs of car production. It is the rising marginal cost that eventually causes average costs to increase.

MARGINAL COST

At each point on the cost function, the *marginal cost* (MC) is the additional cost of producing one more unit of output, which corresponds to the slope of the cost function. If cost increases by ΔC when quantity is increased by ΔQ, the marginal cost can be estimated by:

$$MC = \frac{\Delta C}{\Delta Q}$$

Notice that in Figure 7.7 we calculated *MC* by finding the change in costs, ΔC, for one more car. Sometimes it is more convenient to take a different increase in quantity. If we know that costs rise by $\Delta C = \$12{,}000$ when 5 extra cars are produced, then we could calculate $\Delta C/\Delta Q$, where $\Delta Q = 5$, to get an estimate for *MC* of \$2,400 per car. In general, when the cost function is curved, a smaller ΔQ gives a more accurate estimate.

If your course uses calculus, this Leibniz supplement defines the concepts of average and marginal cost mathematically.

Now look at the shapes of the *AC* and *MC* curves, shown again in Figure 7.8. You can see that at values of Q where the *AC* is greater than the *MC*, the *AC* curve slopes downward, and it is upward-sloping where *AC* is less than *MC*. This is not just a coincidence: it happens whatever the shape of the total cost function. Work with the interactive Figure 7.8 to see why this happens.

DISCUSS 7.2: THE COST FUNCTION FOR APPLE-CINNAMON CHEERIOS

Of course, the cost function in Figure 7.6 is not the only possible shape for a cost function. For Apple-Cinnamon Cheerios, we assumed the unit cost of a pound of cereal was equal to \$2, irrespective of the quantity produced. That is to say, the average cost was constant.

1. Try drawing the cost function for this case.
2. What do the marginal and average costs functions look like?
3. Now suppose that the marginal cost of producing a pound of Cheerios was \$2, whatever the quantity, but there were also some fixed production costs. Trying drawing the total, marginal and average cost curves in this case.

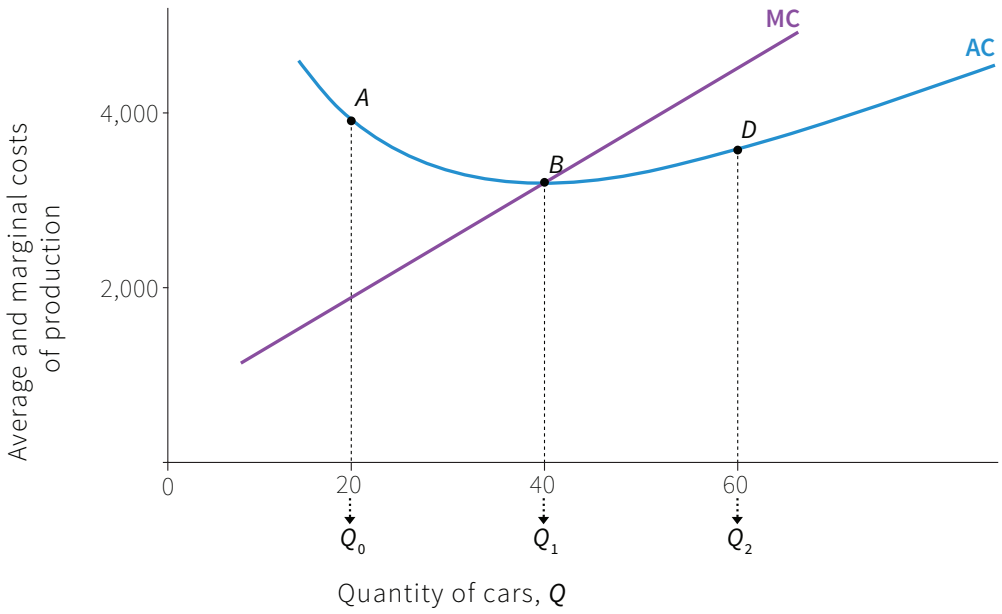


Figure 7.8 *Average and marginal cost curves.*

The diagram shows both the average cost curve and the marginal cost curve.Look at point *A* on the AC curve. When $Q = 20$, the average cost is $4,000, but the marginal cost is only $2,000. So if 21 cars rather than 20 are produced, that will reduce the average cost. Average cost is lower at $Q = 21$. At any point, like point *A*, where $AC > MC$, the average cost will fall if one more car is produced, so the AC curve slopes downward. At point *D* where $Q = 60$, the average cost is $3,600, but the cost of producing a 61st car is $4,600. So the average cost of a car will rise if 61 cars are produced. When $AC < MC$, the average cost curve slopes upward. At point *B*, where the average cost is lowest, the average and marginal costs are equal. The two curves cross. When $AC = MC$, the *AC* curve doesn't slope up or down: it is flat (the slope is zero).

Rajindar and Manjulika Koshal, two economists, studied the cost functions of public universities in the US. They estimated the marginal and average costs of educating graduate and undergraduate students in 171 public universities in the academic year 1990-1, and found decreasing costs. They also found that the universities benefitted from what are termed *economies of scope*: that is, there were cost savings from producing several products—in this case graduate education, undergraduate education, and research—together. If you want to know more about costs, George Stigler, an economist, has written an entertaining discussion of the subject in chapter 7 of this book, which is available online.

DISCUSS 7.3: COST FUNCTIONS FOR UNIVERSITY EDUCATION

Below you can see the average and marginal costs per student for the year 1990-1 that Koshal and Koshal calculated from their research.

	STUDENTS	MC ($)	AC ($)	TOTAL COST ($)
Undergraduates	2,750	7,259	7,659	21,062,250
	5,500	6,548	7,348	40,414,000
	8,250	5,838	7,038	
	11,000	5,125	6,727	73,997,000
	13,750	4,417	6,417	88,233,750
	16,500	3,706	6,106	100,749,000
	STUDENTS	MC ($)	AC ($)	TOTAL COST ($)
Graduates	550	6,541	12,140	6,677,000
	1,100	6,821	9,454	10,399,400
	1,650	7,102	8,672	
	2,200	7,383	8,365	18,403,000
	2,750	7,664	8,249	22,684,750
	3,300	7,945	8,228	27,152,400

Source: *Koshal, Rajindar K., and Manjulika Koshal. 1999. 'Economies of Scale and Scope in Higher Education: A Case of Comprehensive Universities.' Economics of Education Review 18 (2): 269–77.*

1. How do average costs change as the numbers of students rise?
2. Using the data for average costs, find the missing figures in the total cost columns.
3. Plot the marginal and average cost curves for undergraduate education on a graph with costs on the vertical axis, and the number of students on the horizontal axis. On a separate diagram plot the equivalent graphs for graduates.
4. What are the shapes of the total cost functions for undergraduates and graduates? (You could sketch them using what you know about MC and AC; alternatively you could plot them more accurately using the numbers in the Total Cost columns. Hint: they are not straight lines.)
5. What are the main differences between the universities' cost structures for undergraduates and graduates?
6. Can you think of any explanations for the shapes of the graphs you have drawn?

7.4 SETTING THE PRICE: *BEAUTIFUL CARS*

Not all cars are the same. Cars are *differentiated products*. Each make and model is produced by just one firm, and has some unique characteristics of design and performance that differentiate it from the cars made by other firms.

We expect a firm selling a differentiated product to face a downward-sloping demand curve. We have already seen an empirical example in the case of Apple-Cinnamon Cheerios (another differentiated product). If the price of a *Beautiful Car* is high, demand will be low because the only consumers who will buy are those who strongly prefer *Beautiful Cars* to all other makes. As the price falls, more consumers—who might otherwise have purchased a Ford or a Volvo—will be attracted to a *Beautiful Car*.

The demand curve

For any product that consumers might wish to buy, the product demand curve is a relationship that tells you the number of items (the quantity) they will buy at each possible price. For a simple model of the demand for *Beautiful Cars*, imagine that there are 100 potential consumers who would each buy one *Beautiful Car*, today, if the price were low enough.

Each consumer has a *willingness to pay* (WTP) for a *Beautiful Car*, which depends on how much the customer personally values it (given that the customer has the resources to buy it, of course). A consumer will buy a car if the price is less than, or equal to, that person's WTP. Suppose we line up the consumers in order of WTP, highest first, and plot a graph to show how the WTP varies along the line (Figure 7.9). Then if we choose any price, say $P = \$3{,}200$, the graph shows the number of consumers whose WTP is greater than or equal to P. In this case, 60 consumers are willing to pay $3,200 or more, so the demand for cars at a price of $3,200 is 60.

If P is lower, there are a larger number of consumers willing to buy—so the demand is higher. Demand curves are often drawn as straight lines, as in this example, although there is no reason to expect them to be straight in reality: we saw that the demand curve for Apple-Cinnamon Cheerios was not straight. But we do expect demand curves to slope downward: as price rises, the quantity demanded by consumers falls. Conversely, when the available quantity is low, it can be sold at a high price. This relationship between price and quantity is sometimes known as the *Law of Demand*.

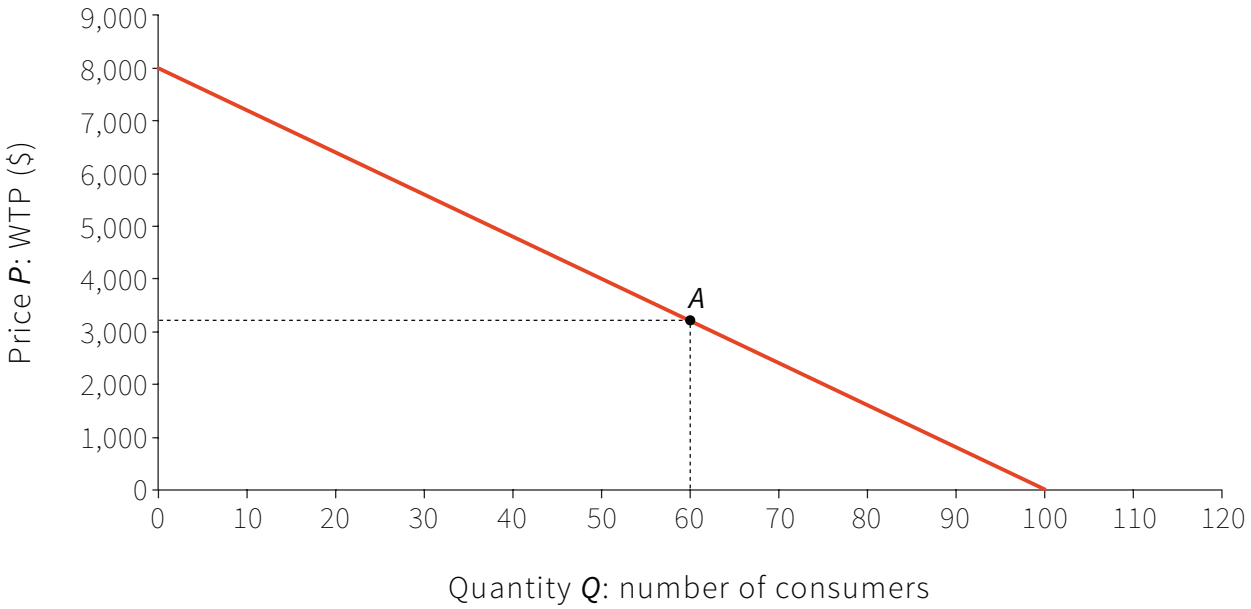


Figure 7.9 *The demand for cars (per day).*

Like the producer of Apple-Cinnamon Cheerios, *Beautiful Cars* will choose the price, P, and the quantity, Q, taking into account its demand curve and its production costs. The demand curve determines the feasible set of combinations of P and Q. To find the profit maximising point, we will draw the isoprofit curves, and look for the point of tangency as before.

The isoprofit curves

The firm's profit is the difference between its revenue (the price multiplied by quantity sold) and its total costs, $C(Q)$:

$$\begin{aligned} profit &= total\ revenue - total\ costs \\ &= PQ - C(Q) \end{aligned}$$

This calculation gives us what is known as the economic profit. Remember that the cost function includes the opportunity cost of capital—the payments that must be made to the owners to induce them to hold shares, which are referred to as normal profits. *Economic profit* is the additional profit above the minimum return required by shareholders.

Equivalently, profit is the number of units of output multiplied by the profit per unit, which is the difference between the price and the average cost:

$$\begin{aligned} profit &= Q\left(P - \frac{C(Q)}{Q}\right) \\ &= Q(P - AC) \end{aligned}$$

For this equation you can see that the shape of the isoprofit curves will depend on the shape of the average cost curve. Remember that for *Beautiful Cars*, the average cost curve slopes downward until Q = *40*, and then upward. Figure 7.10 shows the corresponding isoprofit curves. They look similar to those for Cheerios in Figure 7.3, but there are some differences because the average cost function has a different shape. The lowest (lightest blue) curve shows the zero-economic-profit curve: the combinations of price and quantity for which economic profit is equal to zero—because the price is just equal to the average cost at each quantity.

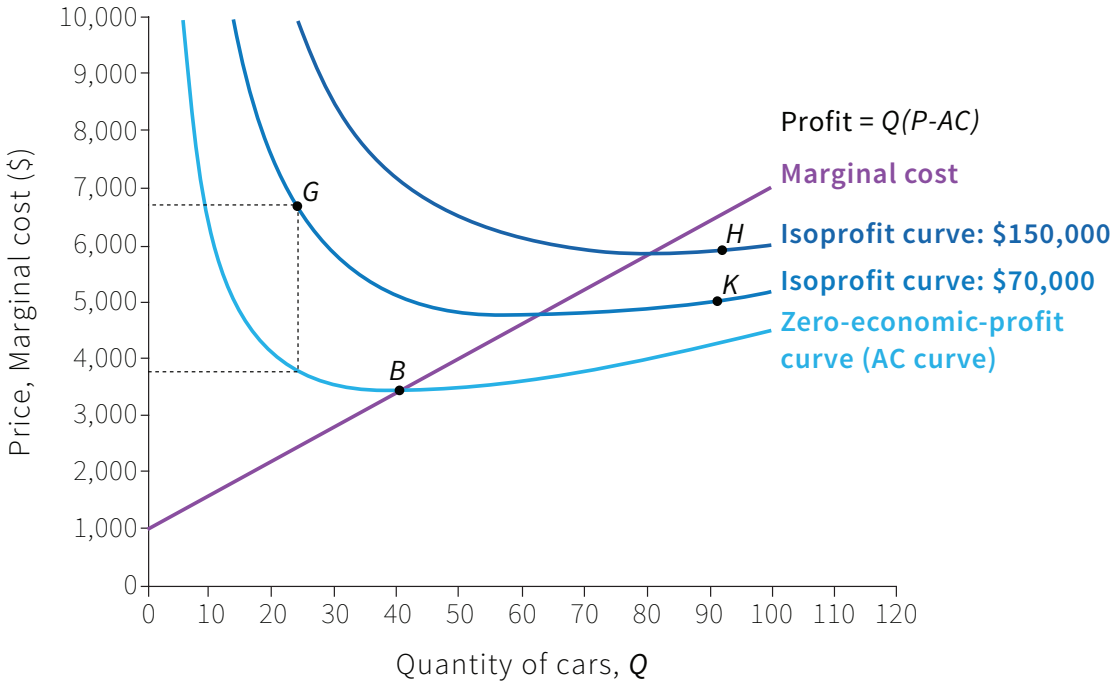


Figure 7.10 *Isoprofit curves for Beautiful Cars.*

The lightest blue curve is the firm's average cost curve. If *P = AC,* the firm's economic profit is zero. So the *AC* curve is also the zero-profit curve: it shows all the combinations of *P* and *Q* which give zero economic profit. *Beautiful Cars* has decreasing *AC* when *Q < 40*, and increasing *AC* when *Q > 40*. When *Q* is low it needs a high price to break even. If *Q = 40* it could break even with a price of $3,400. For *Q > 40*, it would need to raise the price again to avoid a loss. *Beautiful Cars* has increasing marginal cost: the upward-sloping line. Remember that the *AC* curve slopes down if *AC > MC*, and up if *AC < MC*. The two curves cross at *B*, where *AC* is lowest. The darker blue curves show the combinations of *P* and *Q* giving higher levels of profit; so points *G* and *K* give the same profit. At *G* where the firm makes 23 cars, the price is $6,820 but the average cost is $3,777. The firm makes a profit of $3,043 on each car, and its total profit is $70,000. Profit is higher on the curves further up the diagram. Point *H* has the same quantity as *K*, so the average cost is the same, but the price is higher at *H*.

Notice that in Figure 7.10:

- Isoprofit curves slope *downward* at points where $P > MC$.
- Isoprofit curves slope *upward* at points where $P < MC$.

The difference between the price and the marginal cost is called the *profit margin*, and at any point on an isoprofit curve the slope is given by:

$$\text{slope of isoprofit curve} = -\frac{(P - MC)}{Q}$$
$$= -\frac{\text{profit margin}}{\text{quantity}}$$

To understand why, think again about point G in Figure 7.10 at which $Q = 28$, and the price is much higher than the marginal cost. If you:

1. Increase Q by 1
2. Reduce P by $(P - MC)/Q$

Then your profit will stay the same because the extra profit of $(P - MC)$ on car 29 will be offset by a fall in revenue of $(P - MC)$ on the other 28 cars. In the Einstein section we give a fuller explanation, and to find out how to calculate the slope of the isoprofit curve if you are familiar with calculus, see this Leibniz.

DISCUSS 7.4: LOOKING AT ISOPROFIT CURVES

The isoprofit curves for Cheerios are downward-sloping, but for *Beautiful Cars* they slope downward when Q is low and upward when Q is high.

1. What is the reason for the difference?
2. In both cases the higher isoprofit curves get closer to the average cost curve as quantity increases. Why?

Setting price and quantity to maximise profit

In Figure 7.11 we have shown both the demand curve and the isoprofit curves for *Beautiful Cars*. What is the best choice of price and quantity for the manufacturer? The only feasible choices are the points on or below the demand curve, shown by the shaded area on the diagram. To maximise profit the firm should choose the tangency point E, where it reaches the highest possible isoprofit curve.

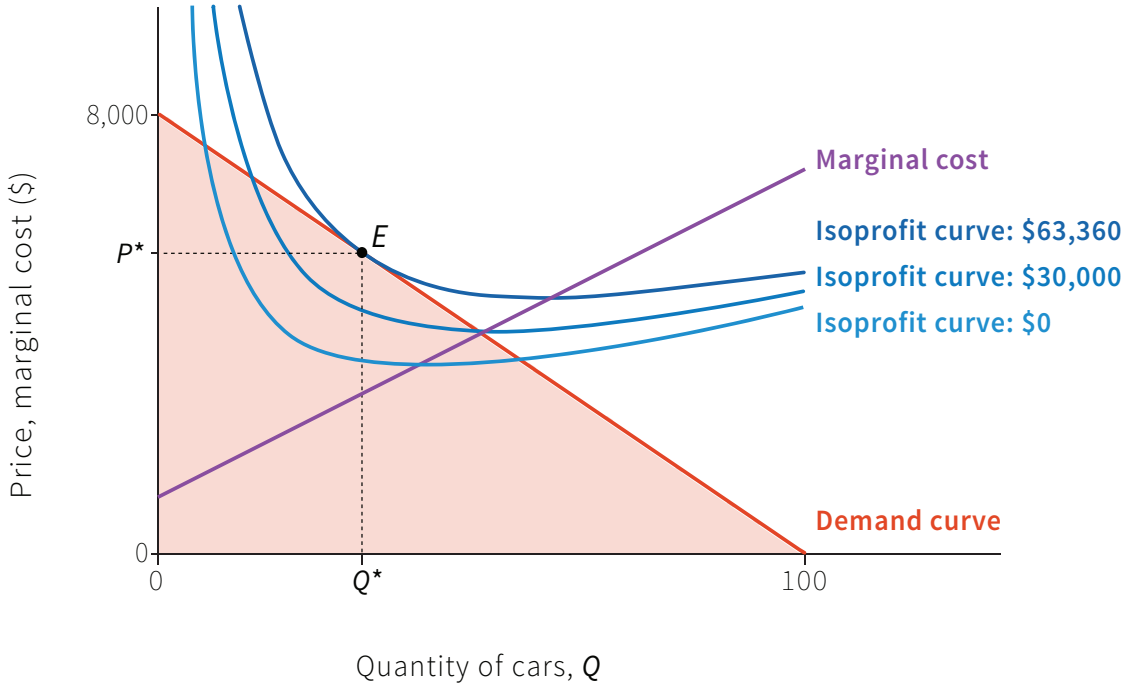


Figure 7.11 *The profit-maximising choice of price and quantity for Beautiful Cars.*

The profit-maximising price and quantity are $P^* = \$5{,}440$ and $Q^* = 32$, and the corresponding profit is $63,360. As in the case of Cheerios, the optimal combination of price and quantity balances the trade-off that the firm would be willing to make between price and quantity (because it doesn't change profit) against the trade-off the firm is constrained to make by the demand curve. This Leibniz supplement shows how to find an equation for the firm's profit-maximising price using calculus.

Figure 7.11 looks very much like earlier figures you have studied, including Alexei's choice of study time, Angela's choice of farming time, and Anil's choice of how to distribute his lottery winnings. It looks the same because it presents the same kind of problem, namely a feasible set of outcomes available to the decision-maker and a set of indifference curves showing his or her evaluation of each of the possible outcomes.

The point chosen by the firm in this case is also familiar:

- The slope of the feasible set (the demand curve in this case) is the *marginal rate of transformation* (MRT) of lower prices into greater quantity sold.
- The slope of the indifference curve is the *marginal rate of substitution* (MRS) in creating profits between selling more and charging more.

So at point E we have the familiar condition balancing condition between the two trade-offs: $MRT = MRS$.

Compared with the multinational giants of the automobile industry, *Beautiful Cars* is a small firm: it chooses to make only 32 cars per day. In terms of its production levels (but not its prices) it is more similar to luxury brands like Aston-Martin, Rolls Royce and Lamborghini, each of which produces fewer than 5,000 cars a year. The size of *Beautiful Cars* is determined partly by its demand function—there are only 100 potential buyers per day, at any price. In the longer term, the firm may be able to increase demand by advertising, to bring its product to the attention of more consumers, and convince them of its desirable qualities. But if it wants to expand production it will also need to look at its cost structure. At present it has rapidly increasing marginal costs, bringing decreasing returns to scale when output per day exceeds 40. With its current premises and equipment it is difficult to produce more than 40 cars. Investment in new equipment may help to reduce its marginal cost, and might make expansion possible.

7.5 ANOTHER WAY TO LOOK AT PROFIT MAXIMISATION: MARGINAL REVENUE AND MARGINAL COST

In the previous section we showed that the profit-maximising choice for *Beautiful Cars* was the point where the demand curve was tangent to an isoprofit curve. To make maximum profit, it should produce $Q = 32$ cars and sell them at a price $P = \$5{,}440$.

We now look at a different method of finding the profit-maximising point—without using isoprofit curves. Instead, we use the marginal revenue curve. Remember that if Q cars are sold at a price P, revenue R is given by $R = P \times Q$. The marginal revenue, MR, is the increase in revenue obtained by increasing the quantity from Q to $Q + 1$.

Figure 7.12a shows you how to calculate the marginal revenue when $Q = 20$. That is, the increase in revenue if the quantity is increased by one unit.

Figure 7.12a shows that the firm's revenue is the area of the rectangle drawn below the demand curve. When Q is increased from 20 to 21, revenue changes for two reasons. An extra car is sold, at the new price, but since the price is the price is lower when $Q = 21$, there is also a loss of \$80 on the other 20 cars. The marginal revenue is the combination of these two changes.

Q = 20	*P* = $ 6,400	*R* = $ 128,000	Gain in revenue (21st car)	$ 6,320
Q = 21	*P* = $ 6,320	*R* = $ 132,720	Loss of revenue ($80 on each of 20 cars)	-$ 1,600
ΔQ = 1	Δ*P* = -$ 80	Δ*R* = $ 4,720	Marginal revenue	$ 4,720

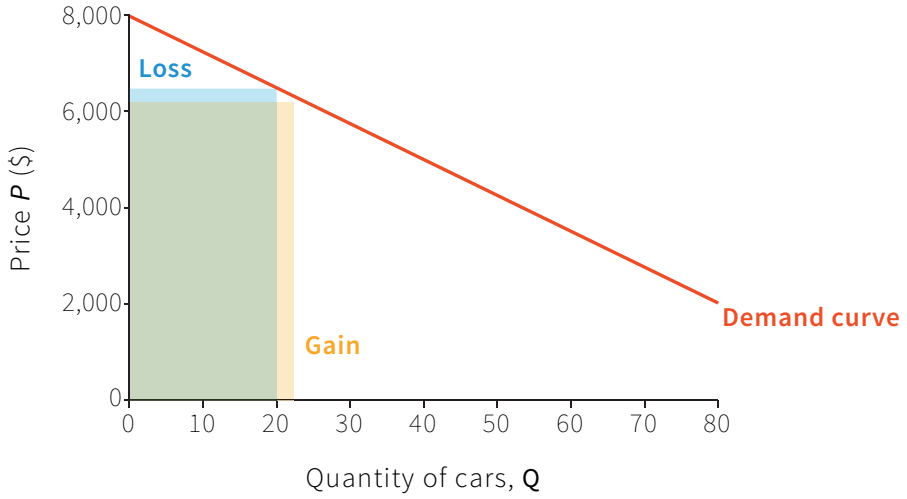


Figure 7.12a *Calculating marginal revenue.*

Figure 7.12b shows how to find the marginal revenue curve, and use it to find the point of maximum profit. The upper panel shows the demand curve, and the middle panel shows both the marginal cost curve, and the marginal revenue curve, which is obtained by calculating *MR* at each point along the demand curve. When *P* is high and Q is low, *MR* is high: the gain from selling one more car is much greater than the total loss on the small number of other cars. As we move down the demand curve *P* falls—so the gain on the last car gets smaller; and Q rises—so the total loss on the other cars is bigger; *MR* falls, and eventually becomes negative.

The marginal revenue curve is usually (although not necessarily) a downward-sloping line. The lower two panels in Figure 7.12b demonstrate that the profit-maximising point is where the MR curve crosses the MC curve. To understand why, remember that profit is the difference between revenue and costs, so for any value of Q, the change in profit if Q was increased by one unit—the marginal profit—would be the difference between the change in revenue, and the change in costs:

$$profit = total\ revenue - total\ costs$$
$$marginal\ profit = MR - MC$$

So:

- *MR* > *MC*: The firm could increase profit by raising Q.
- *MR* < *MC*: The marginal profit is negative. It would be better to decrease Q.

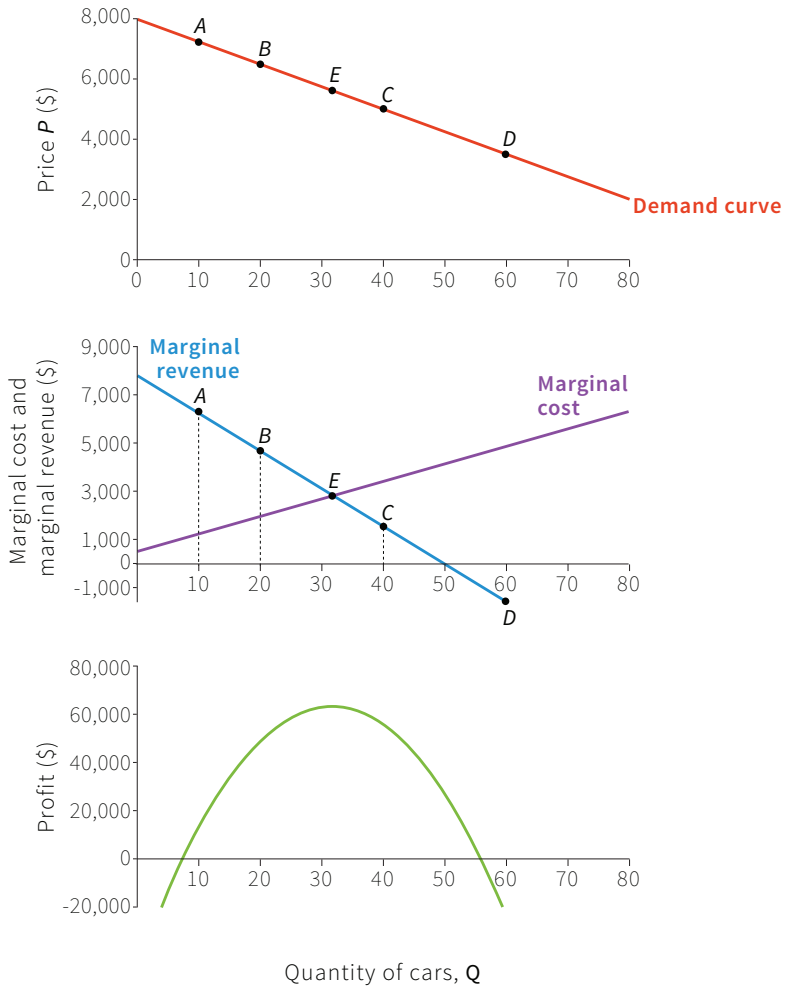


Figure 7.12b *Marginal revenue, marginal cost, and profit.*

The upper panel shows the demand curve and in the middle panel we have drawn the marginal cost curve. At point *A*, where $Q = 10$ and $P = \$7{,}200$, revenue is \$72,000. The marginal revenue at A is the difference between the areas of the two rectangles: $MR = \$6{,}480$. It is plotted in the middle panel. Marginal revenue when $Q = 20$ and $P = \$6{,}400$ is \$4,880. As we move down the demand carve, *P* falls and *MR* falls faster. The gain on the extra car gets smaller, and the loss on the other cars is bigger. At point *D*, the gain on the extra car is outweighed by the loss on the others. Joining the points in the middle panel gives the marginal revenue curve. *MR* and *MC* cross at point *E*, where $Q = 32$. At any value of *Q* below 32, $MR > MC$. The revenue from selling an extra car is greater than the cost of making it, so it would be better to increase production. When $Q > 32$, $MR < MC$: if the firm was producing more than 32 cars it would lose profit if it made an extra car, and it would gain profit if it made fewer cars. In the lower panel we have plotted the firm's profit at each point on the demand curve. You can see that when $Q < 32$, $MR > MC$, and profit increases if *Q* increases. When $Q = 32$, profit is maximised. When $Q > 32$, $MR < MC$, and profit falls if *Q* rises.

You can see how profit changes with Q in the lower panel of 7.12b. Just as marginal cost is the slope of the cost function, marginal profit is the slope of the profit function. In this case:

- When $Q < 32$, $MR > MC$: Marginal profit is positive, so profit increases with Q.
- When $Q > 32$, $MR < MC$: Marginal profit is negative; profit decreases with Q.
- When $Q = 32$, $MR = MC$: Profit reaches a maximum.

7.6 GAINS FROM TRADE

We can analyse the outcome of the economic interactions between consumers and a firm in terms of efficiency and fairness, as we did for Angela and Bruno in Unit 5. We have assumed that rules of the game for allocating Cheerios and cars to consumers are:

1. A firm decides how many items to produce, and sets a price.
2. Then individual consumers decide whether to buy.

These rules reflect typical market institutions for the allocation of consumer goods, although we might imagine alternatives—maybe a group of people who wanted cars could get together to produce a specification, then invite manufacturers to tender for the contract.

In the interactions between a firm like *Beautiful Cars* and its consumers, there are potential gains for both, as long as the firm is able to manufacture a car at a cost less than the value of the car to a consumer. Recall that the demand curve shows the willingness to pay (WTP) of each of the potential consumers. A consumer whose WTP is greater than the price will buy the good and receive a surplus—the value to her of the car is more than she has to pay for it. Similarly the marginal cost curve shows what it costs to make each additional car. (If you start at $Q = 0$, the marginal cost curve shows how much it costs to make the first car, then the second, and so on.) And if the marginal cost is lower than the price, the firm receives a surplus too. Figure 7.13 shows how to find the *total surplus*—recall that this is sometimes called the *gains from trade* or *gains from exchange*—for the firm and its consumers, when *Beautiful Cars* sets the price to maximise its profits.

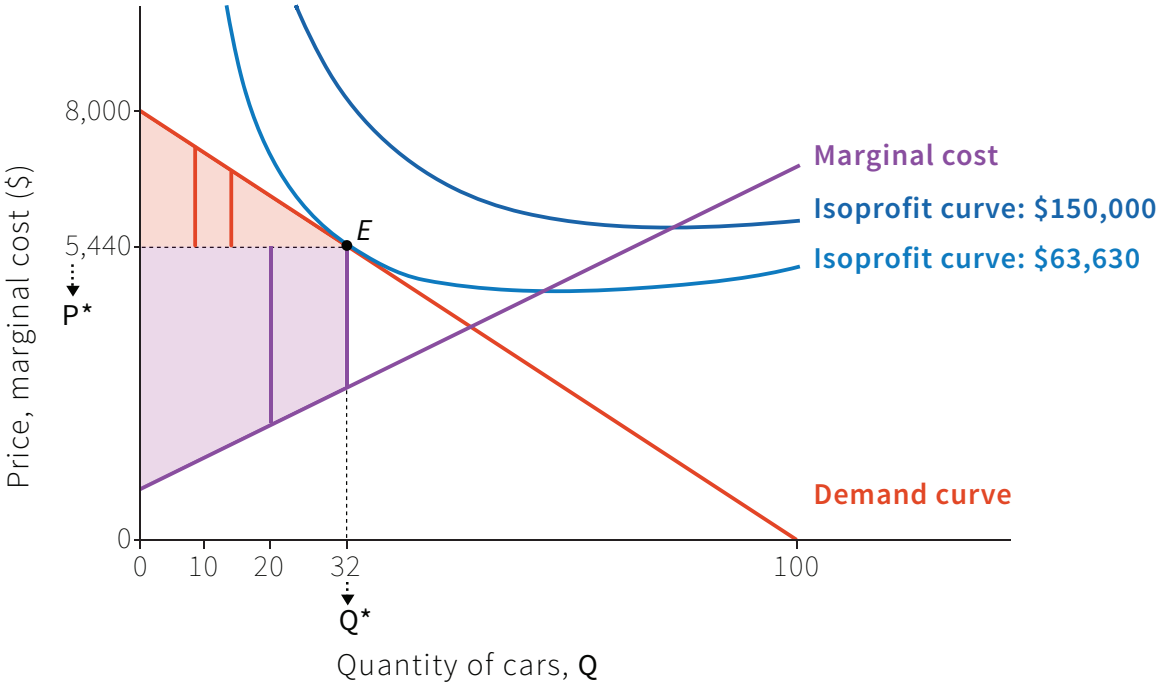


Figure 7.13 *Gains from trade.*

When the firm sets its profit-maximising price $P^* = \$5{,}440$ and sells $Q^* = 32$ cars per day, the 32nd consumer, whose WTP is \$5,440, is just indifferent between buying and not buying a car, so that buyer's surplus is equal to zero. Other buyers were willing to pay more. The 10th consumer, whose WTP is \$7,200, makes a surplus of \$1,760, shown by the short vertical line. The 15th consumer has WTP of \$6,800 and hence a surplus of \$1,360. To find the total surplus obtained by consumers, we add together the surplus of each buyer: this is shown by the shaded triangle between the demand curve and the line where price is P^*. This measure of the consumers' gains from trade is the *consumer surplus*. Similarly, the firm makes a producer surplus on each car sold. The marginal cost of the 20th car is \$2,000. By selling it for \$5,440 the firm gains \$3,440, shown by a vertical line in the diagram between P^* and the marginal cost curve. To find the total producer surplus, we add together the surplus on each car produced: this is the purple-shaded area. The firm obtains a surplus on the marginal car: the 32nd and last car is sold at a price greater than marginal cost.

In Figure 7.13, the shaded area above P^* measures the *consumer surplus*, and the shaded area below P^* is the *producer surplus*. We see from the relative size of the two areas in Figure 7.13 that, in this market, the firm obtains a greater share of the surplus.

CONSUMER SURPLUS, PRODUCER SURPLUS, PROFIT

- The *consumer surplus* is a measure of the benefits of participation in the market for consumers.
- The *producer surplus* is closely related to the firm's profit, but it is not quite the same thing. Producer surplus is the difference between the firm's revenue and the marginal costs of every unit, but it doesn't allow for the fixed costs, incurred even when $Q = 0$.
- The *profit* is the producer surplus minus fixed costs.
- The *total surplus* arising from trade in this market, for the firm and consumers together, is sum of the two areas.

As in the case of the voluntary exchange contracts between Angela and Bruno in Unit 5, both parties gain in the market for *Beautiful Cars*; and the division of the gains is determined by bargaining power. In this case the firm has more power than its consumers because it is the only seller of *Beautiful Cars*. It can set a high price and obtain a high share of the gains, knowing that consumers with high valuations of the car have no alternative but to accept. An individual consumer has no power to bargain for a better deal because the firm has many other potential customers.

Is the allocation of cars in this market Pareto efficient? The answer is no, because there are some consumers who do not purchase cars at the firm's chosen price, who would nevertheless be willing to pay more than it would cost the firm to produce them. In Figure 7.13 we saw that *Beautiful Cars* makes a surplus on the marginal car: that is the 32nd one. The price is greater than the marginal cost. It could produce another car, and sell it to the 33rd consumer at a price lower than $5,440, but higher than the production cost. Both the firm and the 33rd consumer would be better off. In other words, the potential gains from trade in the market for this type of car have not been exhausted at E.

Suppose the firm had chosen instead the point where the marginal cost curve crosses the demand curve. Figure 7.14 demonstrates that this point represents a Pareto efficient outcome, with no further potential gains from trade—if any more cars were produced they would cost more than any of the remaining consumers would pay. The total surplus would have been higher at this point than it is at E.

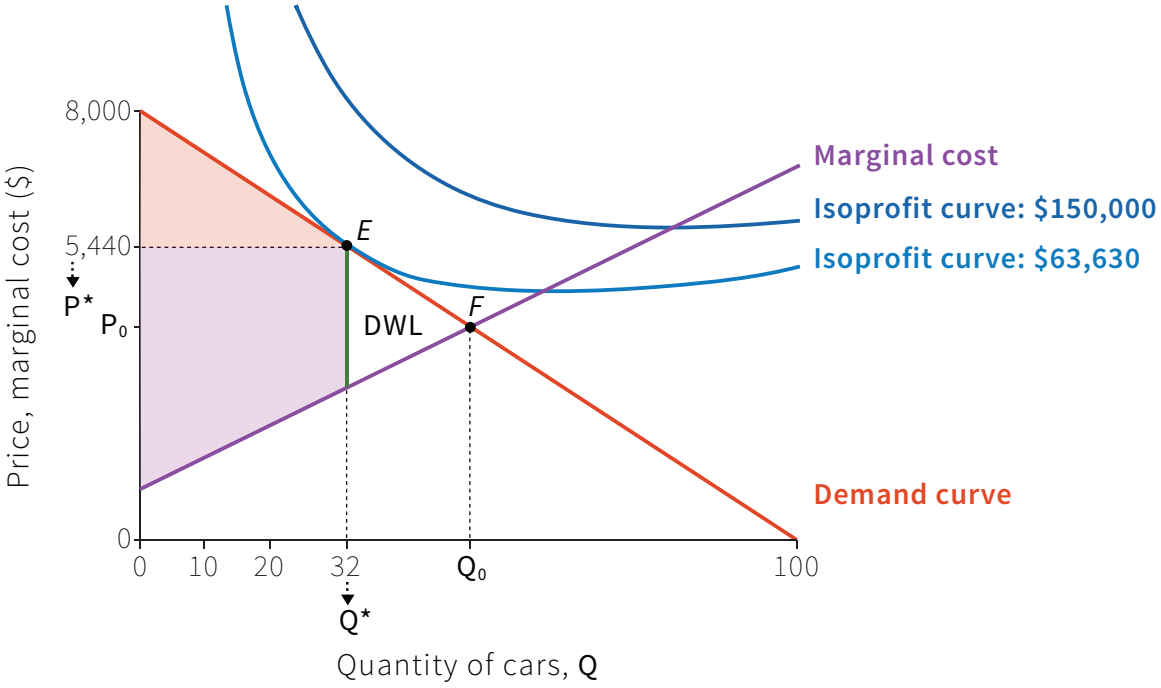


Figure 7.14 *Deadweight loss.*

The total surplus, which we can think of as the pie to be shared between the firm and its customers, would be higher at the Pareto efficient point *F* than at point *E*. Consumer surplus is higher at *F* than *E*, because those who were willing to buy at the higher price would benefit from the decrease in price, and additional consumers would also obtain a surplus. But *Beautiful Cars* will not choose *F*, because producer surplus is lower there (and you can see that it is on a lower isoprofit curve). The loss of potential surplus is known as the *deadweight loss*. On the diagram it is the triangular area between Q = 32, the demand curve, and the marginal cost curve.

It might seem confusing that the firm choose *E* when we said that at this point it would be possible for both the consumers and the firm to be better off. That is true, but only if cars could be sold to other consumers at a lower price than to the first 32 consumers. The firm chooses *E* because that is the best it can do given the rules of the game (setting one price for all consumers). The allocation that results from price-setting by the producer of a differentiated product like *Beautiful Cars* is Pareto inefficient. The firm uses its bargaining power to set a price that is higher than the marginal cost of a car. Furthermore, it keeps the price high by producing a quantity that is too low, relative to the Pareto efficient allocation.

DISCUSS 7.5: PARETO EFFICIENCY

1. What would happen in the market for *Beautiful Cars* if the firm had so much bargaining power that it could charge each consumer, separately, the maximum they would be willing to pay?
2. How many cars would be sold, and what would the producer and consumer surpluses be?
3. Can you think of any examples of goods that are sold in this way?
4. Why is this not common practice?
5. Some firms do charge different prices to different groups of consumers; for example airlines may charge higher fares for last-minute travellers. Why would they do this and what effect would it have on the division of the surplus?
6. Describe some alternative rules of the game that would give consumers more bargaining power.
7. Under these rules, how many cars would be sold?
8. Under these rules, what would the producer and consumer surpluses be?

7.7 THE ELASTICITY OF DEMAND

The firm maximises profit by choosing the point where the slope of the demand curve is equal to the slope of the isoprofit curve. The slope of the demand curve represents the trade-off that the firm is constrained to make between price and quantity.

So the firm's decision depends on how steep the demand curve is: in other words, how much consumers' demand for a good will change if the price changes. The *price elasticity of demand* is a measure of the responsiveness of consumers to a price change: it is defined as the percentage change in demand that would occur in response to a 1% increase in price. For example, suppose that when the price of a product increases by 10%, we observe a 5% fall in the quantity sold. Then we can calculate the elasticity, ε, as follows:

$$\varepsilon = \frac{-\%\ change\ in\ demand}{\%\ change\ in\ price}$$

ε is the greek letter *epsilon*, which is often used for elasticity. Notice that for a demand curve, quantity falls when price increases. Therefore the change in demand is negative if the price change is positive, and vice versa. The minus sign in the formula for the elasticity ensures that we get a positive number as our measure of responsiveness. So in this example we get:

$$\varepsilon = \frac{-(-5)}{10}$$
$$= 0.5$$

The price elasticity of demand is related to the slope of the demand curve: if the demand curve is quite flat, the quantity changes a lot in response to a change in price, so the elasticity is high. Conversely, a steeper demand curve corresponds to a lower elasticity. But they are not the same thing, and it is important to notice that the elasticity changes as we move along the demand curve, *even if the slope doesn't*.

Figure 7.15 shows, again, the demand curve for cars, which has a constant slope: it is a straight line. At every point, if the quantity increases by one ($\Delta Q = 1$), the price falls by \$80 ($\Delta P = -\$80$):

$$\textit{slope of the demand curve} = \frac{\Delta P}{\Delta Q}$$
$$= -80$$

We can use this to find the elasticity at different points on the demand curve. At point A, for example, $Q = 20$ and $P = \$6{,}400$, so if quantity changes by $\Delta Q = 1$, the change in price is $\Delta P = -\$80$ and the elasticity can be calculated:

$$\%\ \textit{change in}\ Q = 100 \times \frac{\Delta Q}{Q}$$
$$= 5\%$$
$$\%\ \textit{change in}\ P = 100 \times \frac{\Delta P}{P}$$
$$= -1.25\%$$

And so:

$$\varepsilon = \frac{-5}{-1.25}$$
$$= 4$$

Figure 7.15 shows that as we move down the demand curve, the same changes in P and Q correspond to a higher percentage change in P and a lower percentage change in Q, so the elasticity falls. We say that demand is *elastic* if the elasticity is higher than 1, and *inelastic* if is it less than 1.

The table also demonstrates that marginal revenue is positive when demand is elastic, and negative when it is inelastic. Why does this happen? When demand is highly elastic, the firm can increase its quantity without much of a fall in price.

So if it produces one extra car, it will gain revenue on the extra car without losing much revenue on the other cars, and its total revenue will rise; in other words, $MR > 0$. Conversely, if demand is inelastic, the firm cannot increase Q without a big drop in P, so $MR < 0$. In The elasticity of demand and the marginal revenue in this unit's Einstein section we demonstrate that this relationship is true for all demand curves.

$$Elasticity = -\frac{\%\ Change\ in\ Q}{\%\ Change\ in\ P}$$

	A	B	C	D
Q	20	40	50	70
P	$ 6,400	$ 4,800	$ 4,000	$ 2,400
ΔQ	1	1	1	1
ΔP	-$ 80	-$ 80	-$ 80	-$ 80
% change in Q	5.00%	2.50%	2.00%	1.43%
% change in P	-1.25%	-1.67%	-2.00%	-3.33%
Elasticity	4.00	1.50	1.00	0.43
MR	$ 4,880	$ 1,680	$ 80	-$ 3,120

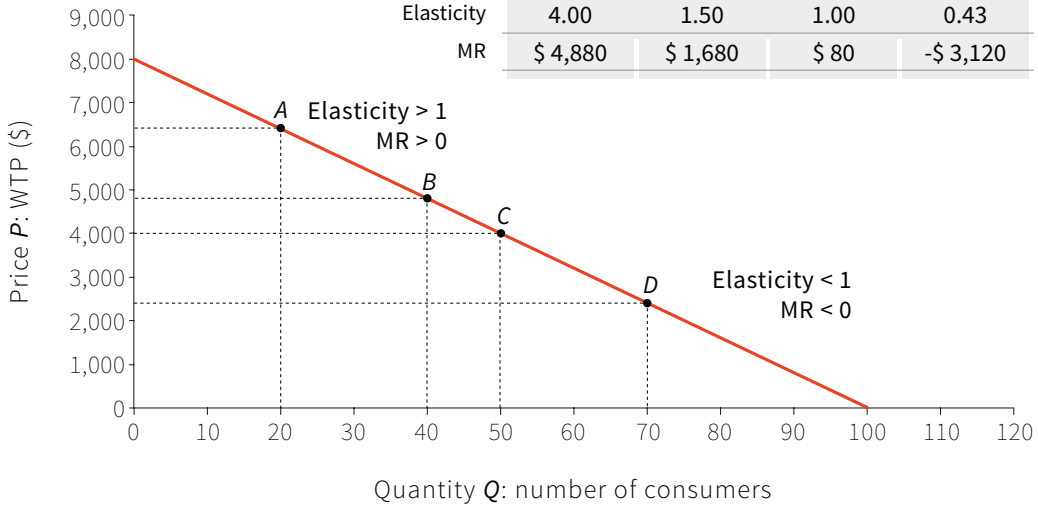


Figure 7.15 *The elasticity of demand for cars.*

How does the elasticity of demand affect a firm's decisions? Remember that the car manufacturer's profit-maximising quantity is Q = 32. You can see in Figure 7.15 that this is on the *elastic* part of the demand curve. The firm would never want to choose a point such as D where the demand curve is inelastic, because the marginal revenue is negative there; it would always be better to decrease the quantity, since that would raise revenue and decrease costs. So the firm will always choose a point where the elasticity is greater than one.

Secondly, the firm's *profit margin*—the difference between the price and the marginal cost of production—is closely related to the elasticity of demand. Figure 7.16 represents a situation of highly elastic demand. The demand curve is quite flat, so small changes in price make a big difference to sales. The profit-maximising choice is point E. You can see that the profit margin is relatively small. This means that the quantity of cars it chooses to make is not far below the Pareto efficient quantity, at point F.

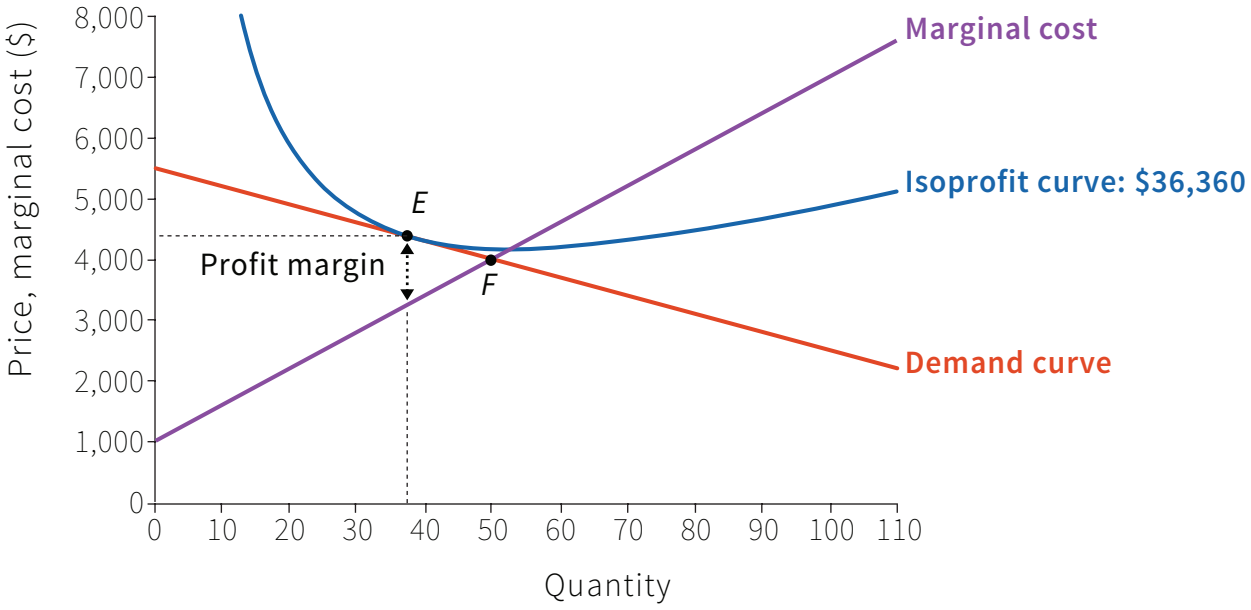


Figure 7.16 *A firm facing highly elastic demand.*

Figure 7.17 shows the decision of a firm with the same costs of car production, but less elastic demand for its product. In this case the profit margin is high, and the quantity is low. When the price is raised, many consumers are still willing to pay. The firm maximises profits by exploiting this situation, obtaining a higher share of the surplus, but the result is that fewer cars are sold and the unexploited gains from trade are high.

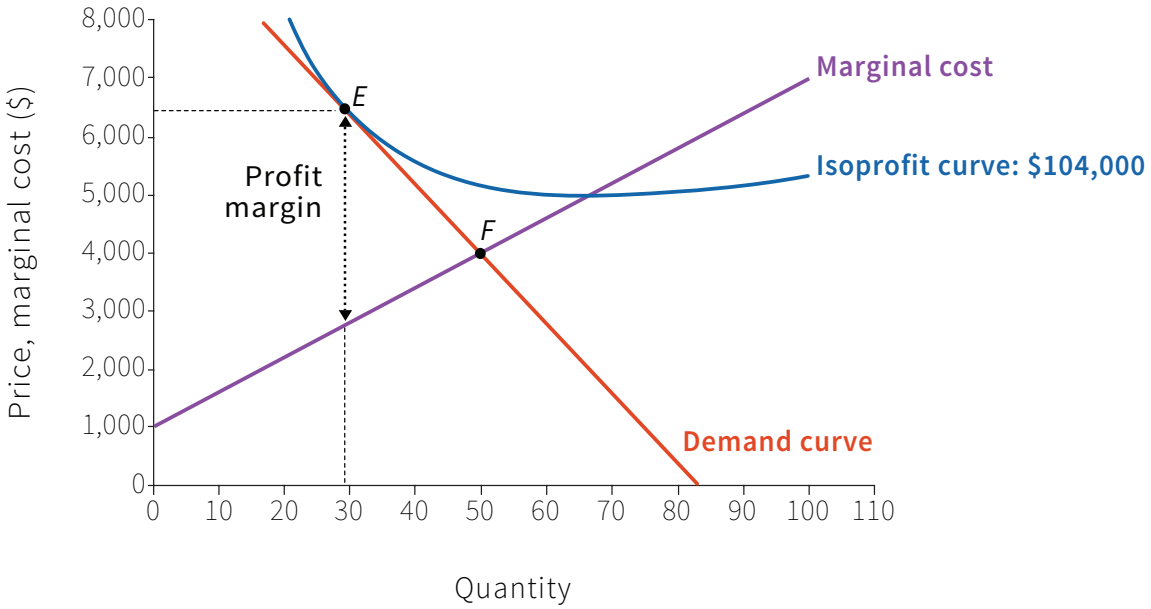


Figure 7.17 *A firm facing less elastic demand.*

These examples illustrate that the lower the elasticity of demand, the more the firm will raise the price above the marginal cost to achieve a high profit margin. When demand elasticity is low, the firm has the power to raise the price without losing many customers. *The size of the markup chosen by the firm* in the Einstein section shows you that the profit margin as a proportion of the price, which we call the *markup*, is inversely proportional to the elasticity of demand. To find out how to calculate elasticities using calculus, see this Leibniz supplement.

7.8 USING DEMAND ELASTICITIES IN GOVERNMENT POLICY

Measuring elasticities of demand is useful to policymakers too. If the government puts a tax on a particular good, the tax will raise the price paid by consumers, so the effect of the tax will depend on the elasticity of demand.

- *If demand is highly elastic, a tax will reduce sales*: That may be what the government intends, for example, governments use taxes on tobacco to discourage smoking because it is harmful to health.
- *But if a tax causes a large fall in sales, it also reduces the potential tax revenue*: This suggests that a government wishing to raise tax revenue should choose to tax products with inelastic demand.

Several countries, including Denmark and France, have recently introduced taxes intended to reduce the consumption of unhealthy food and drink. A 2014 international study found worrying increases in adult and childhood obesity since 1980. In 2013, 37% of men and 38% of women worldwide were overweight or obese. In North America, the figures are 70% and 61%, but the obesity epidemic does not only affect the richest countries; the corresponding rates were 59% and 66% in the Middle East and North Africa.

Matthew Harding and Michael Lovenheim used detailed data on the food purchases of US consumers to estimate elasticities of demand for different types of food, to investigate the effects of food taxes. They divided food products into 33 categories and used a model of consumer decision-making to examine how changes in their prices would change the share of each category in consumers' expenditure on food, and hence the nutritional composition of the diet, taking into account that the change in the price of any product would change the demand for that product and other products too. Figure 7.18 shows the prices and elasticities for some of the categories.

CATEGORY	TYPE	CALORIES PER SERVING	PRICE PER 100G ($)	TYPICAL SPENDING PER WEEK ($)	PRICE ELASTICITY OF DEMAND
1	Fruit and vegetables	660	0.38	2.00	1.128
2	Fruit and vegetables	140	0.36	3.44	0.830
15	Grain, Pasta, Bread	1540	0.38	2.96	0.854
17	Grain, Pasta, Bread	960	0.53	2.64	0.292
28	Snacks, Candy	433	1.13	4.88	0.270
29	Snacks, Candy	1727	0.68	7.60	0.295
30	Milk	2052	0.09	2.32	1.793
31	Milk	874	0.15	1.44	1.972

Figure 7.18 *Price elasticities of demand for different types of food.*

Source: *Extracted from Harding, Matthew, and Machael Lovenheim. 2013. 'The Effect of Prices on Nutrition: Comparing the Impact of Product- and Nutrient-Specific Taxes.' SIEPR Discussion Paper No. 13-023.*

You can see in Figure 7.18 that the demand for lower-calorie milk products is the most price responsive. If their price increased by 10%, the quantity purchased would fall by 19.72%. Demand for snacks and candy is quite inelastic, which suggests that it may be difficult to deter consumers from buying them.

Harding and Lovenheim examined the effects of 20% taxes on sugar, fat and salt. A 20% sugar tax, for example, would increase the price of a product that contains 50% sugar by 10%. A sugar tax was found to have the most positive effect on nutrition. It would reduce sugar consumption by 16%, fat by 12%, salt by 10%, and calorie intake by 19%.

DISCUSS 7.6: ELASTICITY AND EXPENDITURE

Figure 7.18 also shows the spending per week in each category of a US consumer whose total expenditure on food is $80, with typical spending patterns across food categories. Suppose that the price of category 30, high-calorie milk products, increased by 10%:

1. By what percentage would his demand for high-calorie milk products fall?
2. Calculate the quantity he consumes, in grams, before and after the price change.
3. Calculate his total expenditure on high-calorie milk products before and after the price change. You should find that expenditure falls.
4. Now choose a category for which the price elasticity is less than 1, and repeat the calculations. In this case you should find that expenditure rises.

DISCUSS 7.7: FOOD TAXES AND HEALTH

Food taxes intended to shift consumption towards a healthier diet are controversial. Some people think that individuals should make their own choices, and if they prefer unhealthy products, the government should not interfere. Imagine you prefer to eat healthy food, but are sometimes "weak-willed", acting in the short term against your own long-term preferences, and so you make dietary choices you subsequently regret.

Do you think this would justify taxes on sugar, fat, and salt?

7.9 PRICE-SETTING, COMPETITION AND MARKET POWER

Our analysis of the firm's pricing decisions might be applied to any firm producing and selling a product that is in some way different from that of any other firm. In the 19th century the French economist Augustin Cournot carried out a similar analysis using the example of bottled water from "a mineral spring which has just been found to possess salutary properties possessed by no other". Cournot referred to this as a case of *monopoly*—a market in which there is only one seller. He showed, as we have done, that the firm would set a price greater than the marginal production cost.

GREAT ECONOMISTS

AUGUSTIN COURNOT

Augustin Cournot (1801-1877) was a French economist, now most famous for his model of oligopoly (a market with a small number of firms). Cournot's 1838 book *Recherches sur les Principes Mathématiques de la Théorie des Richesses* (Research on the Mathematical Principles of the Theory of Wealth) introduced a new mathematical approach to economics, although he feared it would "draw on me… the condemnation of theorists of repute". Cournot's work influenced other 19th century economists such as Marshall and Walras, and established the basic principles we still use to think about the behaviour of firms. Although he used algebra rather than diagrams, Cournot's analysis of demand and profit maximisation is very similar to ours.

From the previous sections in this unit we know that:

- A firm producing a differentiated good *sets the price above the marginal production cost*.
- The resulting allocation is *Pareto inefficient*—there is a deadweight loss.
- The difference between price and marginal cost—the profit margin—*depends on the elasticity of demand*.

The lower the elasticity of demand, the higher the profit margin and the deadweight loss. So what determines the elasticity of a firm's demand, and why do some firms face more elastic demand than others? To answer this question we need to think again about the decisions made by consumers.

Markets with differentiated products reflect differences in the preferences of consumers. People who want to buy a car are looking for different combinations of characteristics. A consumer's willingness to pay for a particular model will depend not only on its characteristics, but also on the characteristics and prices of similar types of car sold by other firms.

For example, Figure 7.19 shows the purchase prices of a three-door 1.0 litre hatchback in the UK in January 2014 that a consumer could find on a price comparison web site:

		PRICE
FORD FIESTA		£11,917
VAUXHALL CORSA		£11,283
PEUGEOT 208		£10,384
TOYOTA IQ		£11,254

Figure 7.19 *Car purchase prices in the UK (January 2014).*

Source: *Autotrader.com*

Although the four cars are similar in their main characteristics, the website compares them on 75 other features, many of which differ between them.

When consumers are able to choose between several quite similar cars, the demand for each of these cars is likely to be quite elastic. If the price of the Ford Fiesta, for example, were to rise, demand would fall because people would choose to buy one of the other brands instead. Conversely if the price of the Ford Fiesta were to fall, demand would increase because consumers would be attracted away from the other cars. The more similar the other cars are to a Ford Fiesta, the more responsive consumers will be to price differences. Only those with the highest brand loyalty to Ford, and those with a strong preference for a characteristic of the Ford that other cars do not possess, would fail to respond. As we saw in the previous section, highly elastic demand means that the firm will have a relatively low price and profit margin.

In contrast, the manufacturer of a very specialised type of car, quite different from any other brand in the market, faces little competition and hence less elastic demand. It can set a price well above marginal cost without losing customers. Such a firm is earning high rents—economic profits over and above its costs of production—similar to the first firm to introduce a new technology that we saw in Unit 2. The rents arise from its position as the only supplier of this type of car.

So a firm will be in a strong position if it faces little competition from other firms—that is to say, if there are few firms producing close substitutes for its own brand. Then its elasticity of demand will be relatively low. We say that such a firm has *market power*: it will have sufficient bargaining power in its relationship with its customers to set a high price without losing them to competitors.

The problem of market power

This discussion helps to explain why policymakers may be concerned about firms that have few competitors and high market power. They can set high prices, and make high profits, at the expense of consumers. Potential consumer surplus is lost both because few consumers buy, and because those who do buy pay a high price. The owners of the firm benefit, but overall there is a deadweight loss.

A firm selling a niche product catering for the preferences of a small number of consumers (such as a *Beautiful Car* or a luxury brand like a Lamborghini) is unlikely to attract the attention of policymakers, despite the loss of consumer surplus. But if one firm is becoming dominant in a large market, governments may intervene to promote competition. In 2000 the European Commission prevented the proposed merger of Volvo and Scania on the grounds that the merged firm would have a dominant position in the heavy trucks market in Ireland and the Nordic countries, particularly in Sweden where the combined market share of the two firms was 90%. The merged firm would have been almost a monopoly—the extreme case, of a firm that has no competitors at all.

A particular cause for concern is that when there are only a few firms in the market they may form a cartel: a group of firms that collude to keep the price high. By working together, rather than competing with each other, they can increase profits by behaving as a monopoly. A well-known example is OPEC, an association of oil-producing countries. OPEC members jointly agree to set production levels to control the global price of oil. The actions of the OPEC cartel played a major role in sustaining high oil prices at a global level following the sharp increase in oil prices in 1973 and again in 1979. We return to study the effect of the oil price shocks on inflation and unemployment in Unit 14.

While cartels between private firms are illegal in many countries, firms often find ways to cooperate in the setting of prices so as to maximise profits. Policy to limit market power and prevent cartels is known as competition policy, or antitrust policy

in the US. In a famous antitrust case, the US Department of Justice accused Microsoft of behaving anti-competitively by "bundling" its own Internet Explorer web-browser with its Windows operating system.

As the Microsoft example illustrates, dominant firms may exploit their position in ways other than setting high prices. In the 1920s an international group of companies making electric light bulbs, including Philips, Osram and General Electric, formed a cartel that agreed a policy of "planned obsolescence": to reduce the lifetime of their bulbs to 1,000 hours, so that consumers would have to replace them more quickly. The growth of Walmart has been controversial, even though the shops promise "always low prices" for consumers; some people accuse Walmart of using its power in ways they consider unfair to reduce wages in the area around its stores, or to drive smaller retailers out of the market, or to reduce the profits of its wholesale suppliers to unsustainable levels. This paper examines the economic basis for these claims.

DISCUSS 7.8: MULTINATIONALS OR INDEPENDENT RETAILERS?

Imagine that you are a local politician in a town where a multinational retailer is planning to build a new superstore. A local campaign is protesting that it will drive small independent retailers out of business, and thereby reduce consumer choice and change the character of the area. Supporters of the plan counter that this will only happen if consumers prefer the supermarket.

Which side are you on?

Competition policy is not a solution in all cases. In domestic utilities such as water, electricity and gas, there are high fixed costs of providing the supply network, irrespective of the quantity demanded by consumers. Utilities typically have increasing returns to scale. The average cost of producing a unit of water, electricity or gas will be very high unless the firm operates at a large scale. If a single firm can supply the whole market at lower average cost than two firms, the industry is said to be a natural monopoly.

In the case of a natural monopoly, a policymaker may choose to regulate the firm's activities, aiming to increase consumer surplus by limiting the firm's discretion over prices. An alternative is public ownership. The majority of water supply companies around the world are owned by the public sector, although in England and Wales in 1989, and in Chile in the 1990s, the entire water industry was privatised and is regulated by a public sector agency.

7.10 PRODUCT SELECTION, INNOVATION AND ADVERTISING

The profits that a firm can achieve depend on the demand curve for its product, which in turn depends on the preferences of consumers and the competition from other firms. But the firm may be able to move the demand curve to increase profits by its selection of products, or through advertising.

When deciding what goods to produce, the firm would ideally like to find a product that is both attractive to consumers and has different characteristics from the products sold by other firms. In this case demand would be high—many consumers would wish to buy it at each price—and the elasticity low. Of course, this is not likely to be easy: a firm wishing to make a new breakfast cereal, or type of car, knows that there are many brands on the market already. But technological innovation may provide opportunities to get ahead of competitors. For some years after Toyota developed the first mass-produced hybrid car, the Prius, in 1997, there were very few comparable cars available. Toyota effectively monopolised the hybrid market. By 2013 there were several competing brands, but the Prius remained the market leader, with more than 50% of hybrid sales.

If a firm has invented or created a new product, it may be able to prevent competition altogether by claiming exclusive rights to produce it, using patent or copyright laws. Ironically, in the 1970s a company called Parker Brothers spent years fighting in court to protect a monopoly that they had on a profitable board game—called *Monopoly*. This kind of legal protection of monopoly may help to provide incentives for research and development of new products, but at the same time limits the gains from trade. In Unit 20 we analyse intellectual property rights in more detail.

Advertising is another strategy firms can use to influence demand: it is widely used by both car manufacturers and breakfast cereal producers. When products are differentiated, the firm can use advertising to inform consumers about the existence and the characteristics of its product, to attract them away from its competitors, and to create brand loyalty.

According to Schonfeld and Associates, a firm of market analysts, advertising on breakfast cereals in the US is about 5.5% of total sales revenue—about 3.5 times higher than the average for manufactured products. The data in Figure 7.20 is for the highest-selling 35 breakfast cereal brands sold in the Chicago area in 1991 and 1992. The graph shows the relationship between market share and quarterly expenditure on advertising. If you investigated the breakfast cereals market more closely, you would see that market share is not closely related to price. But it is clear from Figure 7.20 that the brands with the highest share are also the ones that spend the most on advertising. Matthew Shum, an economist, analysed cereal purchases in Chicago using this data, and showed that advertising was more effective than price discounts

in stimulating demand for a brand. Since the most well-known brands were also the ones spending most on advertising, he concluded that its main function was not to inform consumers about the product, but rather to increase brand loyalty, and encourage consumers of other cereals to switch.

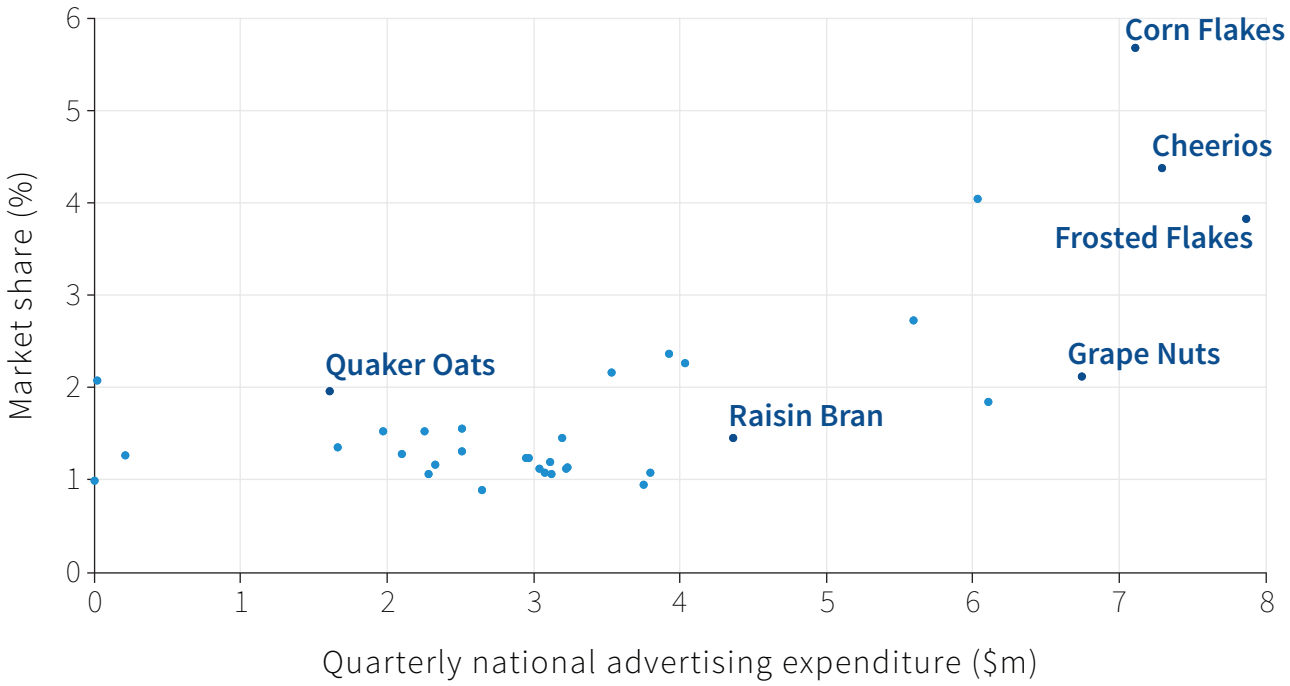


Figure 7.20 *Advertising expenditure and market share of breakfast cereals in Chicago (1991-92).*

Source: *Figure 1 in Shum, Matthew. 2004. 'Does Advertising Overcome Brand Loyalty? Evidence from the Breakfast-Cereals Market.' Journal of Economics & Management Strategy 13 (2): 241–72.*

7.11 CONCLUSION

We have studied how firms producing differentiated products choose the price, and the quantity of output to produce, to maximise their profit. These decisions depend on the demand curve for the product—especially the elasticity of demand—and the cost structure for producing it.

Since the demand curve and cost structure limit the profits that a firm can make, it will be continually looking for ways to influence them. By innovating to develop products that appeal to consumers and are differentiated from others available,

and advertising widely and effectively, it can boost demand and lower the demand elasticity. Innovation in the production process, or investment to enable it to expand the scale of production, may enable it to reduce its average costs.

CONSTRAINED OPTIMISATION PROBLEMS

A decision-maker chooses the values of one or more variables (such as P and Q)...

- ... to achieve an objective (such as maximising profit)
- ... subject to a constraint that determines the feasible set (such as the demand curve)

While firms prefer to operate in markets where there is little direct competition, giving them the power to set high prices and raise profits, high prices lower sales, reducing consumer surplus and causing deadweight loss. Increasing returns to scale may mean that it is efficient for firms to operate at a large scale, but policymakers will be concerned about market power and deadweight loss when large firms achieve a dominant position in a market. Competition policy and regulation are tools they can use to limit the exercise of market power.

In the model of firm behaviour developed in this unit, the firm chooses its price and quantity to maximise its profits, from the feasible set determined by the demand curve. As you have seen, this problem has a similar structure to the one faced by Angela in Unit 3 and Unit 5, who wanted to choose her consumption and working time to maximise her utility within the feasible set determined by her production possibilities.

It is also similar to the problem faced by Alexei in Unit 3, who faced a budget constraint, which in turn defined the feasible set.

In economics (and in mathematics) such problems are known as constrained optimisation problems.

In Unit 6, the firm was the decision-maker choosing the wage to maximise its profits. The feasible set was Maria's best response function. You will see many more examples of constrained optimisation in your study of economics.

CONCEPTS INTRODUCED IN UNIT 7

Before you move on, review these definitions:

- *Differentiated product*
- *Economies of scale*
- *Cost function*
- *Willingness to pay*
- *Demand curve*
- *Price-setting*
- *Consumer surplus*
- *Producer surplus*
- *Deadweight loss*
- *Elasticity of demand*
- *Profit margin*

Key points in Unit 7

The product demand curve

The product demand curve tells you how many units consumers will buy at each price.

The firm's marginal cost

The firm's marginal cost is the addition to total cost of making one extra unit of output.

The profit-maximising point

In markets where products are differentiated, each firm chooses its price and quantity from the feasible set given the demand for its own brand; the profit-maximising point is where the demand curve touches the highest isoprofit curve.

Deadweight loss

In markets where products are differentiated, the firm sets a price greater than its marginal cost, which means that all the potential gains from trade are not realised.

The impact of demand elasticity on price

In markets where products are differentiated the lower the elasticity of demand (steeper demand curve), the higher the firm's price relative to its marginal cost, the higher its profit margin, and the higher the deadweight loss.

7.12 EINSTEIN

The size and cost of a pipe

We can use simple mathematics to work out how much the cost increases when the size of the pipe doubles. The formula for the area of a circle is:

$$area\ of\ circle = \pi \times (radius\ of\ circle)^2$$

Let us assume the area of the pipe was originally 10cm2, and then it was doubled in size to 20cm2. We can use the equation above to find the radius of the pipe in each case.

When the area of the pipe is 10:

$$radius = \sqrt{\frac{10}{\pi}} = 1.78cm$$

When the area of the pipe is 20:

$$radius = \sqrt{\frac{20}{\pi}} = 2.52cm$$

We can now work out the circumference of the pipe, which tells us the cost of pipe in each case. The formula for the circumference of a circle is:

$$circumference = 2 \times \pi \times radius\ of\ circle$$

When the area of the pipe is 10:

$$circumference = 2 \times \pi \times 1.78 = 11.18cm$$

When the area of the pipe is 20:

$$circumference = 2 \times \pi \times 2.52 = 15.83cm$$

The pipe has doubled in capacity, but the circumference, and hence the cost of the pipe, has only increased by a factor of:

$$\frac{15.83}{11.18} = 1.42$$

We can clearly see that the firm has benefitted from economies of scale.

Diseconomies of scale: CORE's Dilbert law of firm hierarchy

If every ten employees at a lower level must have a supervisor at a higher level, then a firm that has 10^x production workers (the bottom of the ladder) will have x levels of management, 10^{x-1} supervisors at the lowest level, 10^{x-2} at the second lowest level, and so on.
— CORE's Dilbert law of firm hierarchy

A firm with a million (106) production workers will thus have 100,000 ($10^5 = 10^{6-1}$) lowest level supervisors. Dilbert did not invent the law, he is too closely watched by his supervisor to have time for that. The CORE team did.

Calculate the slope of the isoprofit curve

An isoprofit curve shows all the combinations of P and Q that give the same level of profit. The firm only cares about the level of profit, so it is indifferent between these combinations. Suppose that a firm is currently selling Q cars at a price P, with a marginal cost of MC, and that $P > MC$.

If the firm increased its quantity by 1, it would receive extra profit of $(P - MC)$. If at the same time it reduced its price by:

its revenue on the other Q cars would fall by $(P - MC)$. So profit would stay the same; if it made these changes, the firm would stay on the same isoprofit curve.

To summarise, if:

$$\Delta Q = 1$$
$$\Delta P = \frac{-(P - MC)}{Q}$$

change in profit = 0.

So the slope of the isoprofit curve is:

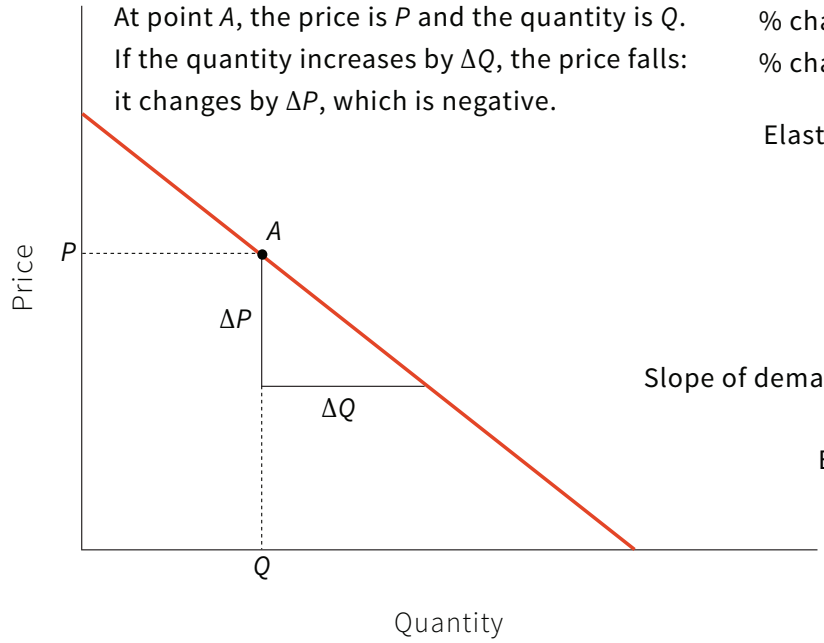
$$\frac{\Delta P}{\Delta Q} = \frac{-(P - MC)}{Q}$$

So when $P > MC$, the isoproft curve slopes downward. The same calculation works when $P < MC$. In this case an increase in price is required to keep profit constant when quantity rises by 1. The isoprofit curve slopes upward.

The elasticity of demand and the marginal revenue

The diagram shows how to obtain a general formula for the elasticity at a point (Q, P) on the demand curve.

It also shows how the elasticity is related to the slope of the demand curve. A flatter demand curve has a lower slope, so a higher elasticity.



$$\%\text{ change in } P = 100 \times \Delta P/P$$
$$\%\text{ change in } Q = 100 \times \Delta Q/Q$$
$$\text{Elasticity at } A = -\frac{\%\text{ change in } Q}{\%\text{ change in } P}$$
$$= -\frac{\Delta Q / Q}{\Delta P / P}$$
$$= -\frac{P}{Q} \times \frac{\Delta Q}{\Delta P}$$
$$\text{Slope of demand curve} = \frac{\Delta P}{\Delta Q}$$
$$\text{Elasticity} = -\frac{P}{Q} \times \frac{1}{\text{slope}}$$

Suppose that the demand curve is *elastic* at A. Then the elasticity is greater than one.

Multiplying by $-Q\Delta P$ (which is positive):

$$P\Delta Q > -Q\Delta P$$

and rearranging, we get:

$$P\Delta Q + Q\Delta P > 0$$

Consider the special case when $\Delta Q = 1$. The inequality becomes:

$$P + Q\Delta P > 0$$

Now remember that the marginal revenue at point A is the change in revenue when Q is increased by one unit, consisting of the gain in revenue on the extra unit, which is P, and the loss on the other units, which is $Q\Delta P$. So this inequality tells us that the marginal revenue is positive.

We have shown that if the demand curve is elastic, $MR > 0$. Similarly, if the demand curve is inelastic, $MR < 0$.

The size of the markup chosen by the firm

We can find a formula that shows that the markup is high when the elasticity of demand is low.

From Figure 7.8 we can see that at the point chosen by the firm, the slope of the isoprofit curve is equal to the slope of the demand curve. We know that the slope of the demand curve is related to the price elasticity of demand:

$$\varepsilon = \frac{P}{Q} \times \frac{1}{slope}$$

Rearranging this formula:

$$slope\ of\ demand\ curve = \frac{P}{Q} \times \frac{1}{elasticity}$$

We also know from section 7.4:

$$slope\ of\ isoprofit\ curve = \frac{-(P - MC)}{Q}$$

When the two slopes are equal:

$$\frac{(P - MC)}{Q} = \frac{P}{Q} \times \frac{1}{elasticity}$$

Rearranging this gives us:

$$\frac{(P - MC)}{P} = \frac{1}{elasticity}$$

The left-hand side is the profit margin as a proportion of the price, which is a measure of the markup. Therefore:

- *The firm's markup is inversely proportional to the elasticity of demand*

7.13 READ MORE

Bibliography

1. Basker, Emek. 2007. 'The Causes and Consequences of Wal-Mart's Growth.' *Journal of Economic Perspectives* 21 (3): 177–98.
2. Cournot, Augustin, and Irving Fischer. (1838) 1971. *Researches into the Mathematical Principles of the Theory of Wealth*. New York, NY: A. M. Kelley.
3. Evans, Heberton G. 1967. 'The Law of Demand--The Roles of Gregory King and Charles Davenant.' *The Quarterly Journal of Economics* 81 (3).
4. Gilbert, Richard J., and Michael L. Katz. 2001. 'An Economist's Guide to US v. Microsoft.' *Journal of Economic Perspectives* 15 (2): 25–44.
5. Harding, Matthew, and Machael Lovenheim. 2013. 'The Effect of Prices on Nutrition: Comparing the Impact of Product- and Nutrient-Specific Taxes.' *SIEPR Discussion Paper No. 13-023*.
6. Hausman, Jerry A. 1996. 'Valuation of New Goods under Perfect and Imperfect Competition.' In *The Economics of New Goods*, by Robert J. Gordon and Timothy F. Bresnahan, 207–48. Chicago, IL: University of Chicago Press.
7. Kay, John. 2015. 'The Structure of Strategy (reprinted from Business Strategy Review 1993).' *Johnkay.com*. Accessed July.
8. Koshal, Rajindar K., and Manjulika Koshal. 1999. 'Economies of Scale and Scope in Higher Education: A Case of Comprehensive Universities.'*Economics of Education Review* 18 (2): 269–77.
9. Luttmer, Erzo G. J. 2011. 'On the Mechanics of Firm Growth.' *The Review of Economic Studies* 78 (3): 1042–68.
10. Schumacher, Ernst F. 1973. *Small Is Beautiful: Economics as If People Mattered*. New York, NY: HarperCollins.
11. Shum, Matthew. 2004. 'Does Advertising Overcome Brand Loyalty? Evidence from the Breakfast-Cereals Market.' *Journal of Economics & Management Strategy* 13 (2): 241–72.
12. Statista. 2011. 'Willingness to Pay for a Flight in Space.' October 20.
13. Stigler, George J. 1987. *The Theory of Price*. New York, NY: Collier Macmillan.
14. *The Economist*. 2008. 'Economies of Scale and Scope', October 20.
15. *The Huffington Post*. 2014. 'There's An Easy Way To Fight Obesity, But Conservatives Will HATE It', January 23.
16. Vickers, John. 1996. 'Market Power and Inefficiency: A Contracts Perspective.' *Oxford Review of Economic Policy* 12 (4): 11–26.
17. Wong, Clement. 2012. 'Planned Obsolescence: The Light Bulb Conspiracy.' Economics Student Society of Australia (ESSA). September 12.

8

SUPPLY AND DEMAND: PRICE-TAKING AND COMPETITIVE MARKETS

Photo: *Abhijit Kar Gupta*

HOW MARKETS OPERATE WHEN ALL BUYERS AND SELLERS ARE PRICE-TAKERS

- Competition can constrain buyers and sellers to be price-takers
- The interaction of supply and demand determines a market equilibrium where both buyers and sellers are price-takers, called a competitive equilibrium
- Price and quantity in the market equilibrium change in response to supply and demand shocks, in the short run and the long run
- Price-taking ensures that all gains from trade in the market are exhausted at a competitive equilibrium
- The model of perfect competition describes idealised conditions under which all buyers and sellers are price-takers
- Real world markets are not typically perfectly competitive, but some policy problems can be analysed using the demand and supply model
- Similarities and differences between price-taking and price-setting firms

Institute for New Economic Thinking
UCL SciencesPo Azim Premji University
FRIENDS PROVIDENT Foundation

Students of American history learn that the defeat of the southern Confederate states in the American civil war ended slavery in the production of cotton and other crops in that region. There is also an economics lesson in this story.

At the war's outbreak, on 12 April 1861, President Abraham Lincoln ordered the US Navy to blockade the ports of the Confederate states. These states had declared themselves independent of the US so as to preserve the institution of slavery.

As a result of the naval blockade, the export of US-grown raw cotton to the textile mills of Lancashire in England came to a virtual halt, eliminating three-quarters of the supply of this critical raw material. Sailing at night, a few blockade-running ships evaded Lincoln's patrols; but 1,500 were destroyed or captured.

We will see in this unit that the market price of a good such as cotton is determined by the interaction of supply and demand. In the case of raw cotton, the tiny quantities reaching England through the blockade were a dramatic reduction in supply. There was large *excess demand*—that is to say, at the prevailing price, the quantity of raw cotton demanded exceeded the available supply. As a result some sellers realised they could profit by raising the price. Eventually, cotton sold at prices six times higher than before the war, keeping the lucky blockade-runners in business.

Mill owners responded. For them, the price rise was an increase in their costs. They cut production to half the pre-war level, throwing hundreds of thousands of people out of work. Some firms failed and left the industry due to the reduction in their profits. Mill owners looked to India to find an alternative to US cotton, greatly increasing the demand for cotton there. The excess demand in the markets for Indian cotton gave some sellers an opportunity to profit by raising prices, resulting in increases in prices of Indian cotton, which quickly rose almost to match the price of US cotton.

Responding to the higher income now obtainable from growing cotton, Indian farmers abandoned other crops and grew cotton instead. The same occurred wherever cotton could be grown, including Brazil. In Egypt, farmers who rushed to expand production of cotton in response to the higher prices began employing slaves, captured (like the American slaves that Lincoln was fighting to free), in sub-Saharan Africa.

There was a problem. The only source of cotton that could come close to making up the shortfall from the US was in India. But Indian cotton differed from American cotton, and required an entirely different kind of processing. Within months of the shift to Indian cotton new machinery was developed to process it.

As the demand for this new equipment soared we know that Dobson and Barlow, a large firm making textile machinery, saw its profits take off. We know this because detailed sales records for this firm have survived. It responded with increased production of these new machines and other equipment. No mill could afford to be

left behind in the rush to retool, because if it didn't, it could not use the new raw material. The result was "such an extensive investment of capital that it amounted almost to the creation of a new industry."

The lesson for economists: Lincoln ordered the blockade. But in what followed, the farmers and sellers who increased the price of cotton were not responding to orders. Neither were the mill owners who cut back the output of textiles and laid off the mill workers, nor were the mill owners desperately searching for new sources of raw material. By ordering new machinery, the mill owners set off a boom in investment and new jobs.

All of these decisions took place over a matter of months, by millions of people, most of whom were total strangers to one another, each seeking to make the best of a totally new economic situation. American cotton was now scarcer, and people responded, from the cotton fields of Maharashtra in India to the Nile delta, to Brazil and the Lancashire mills.

To understand how the change in the price of cotton transformed the world cotton and textile production system, think about the prices determined by markets as messages. The increase in the price of US cotton shouted: "find other sources, and find new technologies appropriate for their use." Similarly, when the price of petrol rises the message to the car driver is: "take the train". It is passed on to the railway operator: "there are profits to be made by running more train services". When the price of electricity goes up, the firm or the family is being told: "think about installing photo-voltaic cells on the roof."

In many cases—like the chain of events that began at Lincoln's desk on 12 April 1861—the messages make sense not only for individual firms and families but also for society: if something has become more expensive then it is likely that more people are demanding it, or the cost of producing it has risen, or both. By finding an alternative, the individual is saving money and in so doing conserving society's resources. This is because, under some conditions, prices provide an accurate measure of the scarcity of a good or service.

In planned economies, which operated in the Soviet Union and other central and eastern European countries before the 1990s (we discussed them in Unit 1), messages about how things would be produced are sent deliberately by government experts. They decide what is produced and at what price it is sold. The same is true, as we saw in Unit 6, in large firms like General Motors, where managers (and not prices) determine who does what.

The amazing thing about prices determined by markets is that individuals do not send the messages; they result from the anonymous interaction of sometimes millions of people, governed by supply and demand. And when conditions change—a cheaper way of producing bread, for example—nobody has to change the message ("put bread instead of potatoes on the table tonight"). A price change results from a change in firms' costs. The reduced price of bread says it all.

8.1 BUYING AND SELLING: DEMAND AND SUPPLY

In Unit 7 we considered the case of a good produced and sold by just one firm. There was one seller and many buyers in the market for that product. In this unit we look at markets where many buyers and sellers interact, and show how the competitive market price is determined by both the preferences of consumers and the costs of suppliers. When there are many firms producing the same product, each firm's decisions are affected by the behaviour of competing firms, as well as consumers.

For a simple model of a market with many buyers and sellers, think about the potential for trade in second-hand copies of a recommended textbook for a university economics course. Demand for the book comes from students who are about to begin the course, and they will differ in their *willingness to pay* (WTP). No one will pay more than the price of a new copy in the campus bookshop. Below that, students' WTP may depend on how hard they work, how important they think the book is, and on the available resources for buying books. Online auctions, such as on eBay, help reveal our willingness to pay, because the price is flexible.

Figure 8.1 shows the demand curve. As in Unit 7, we line up all the consumers in order of willingness to pay, highest first. The first student is willing to pay \$20, the 20th \$10, and so on. For any price, P, the graph tells you how many students would be willing to buy: it is the number whose WTP is at or above P.

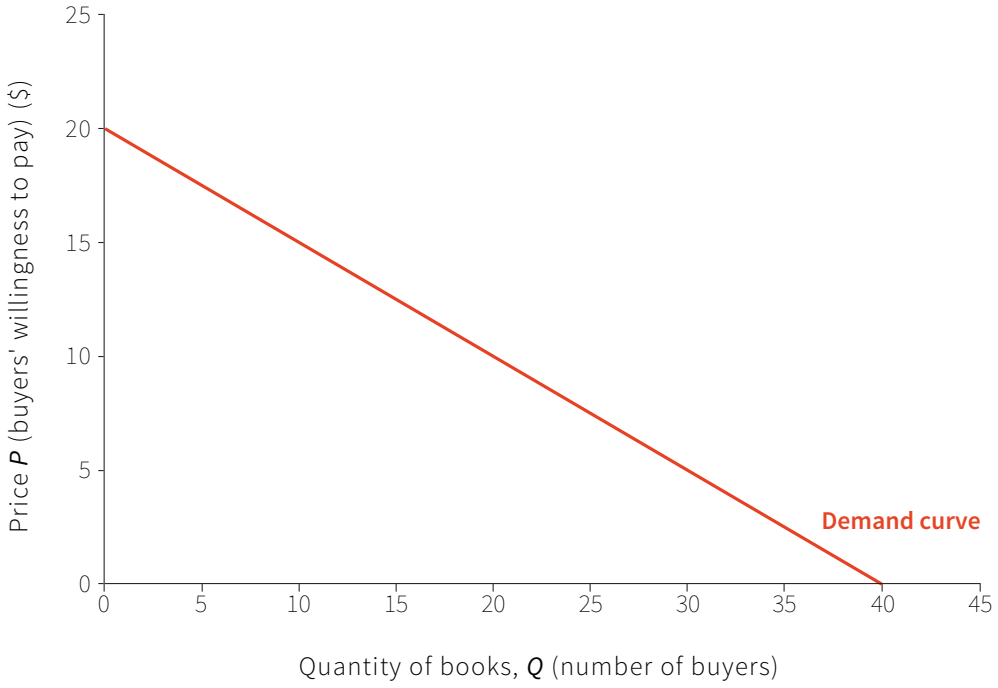


Figure 8.1 *Market demand curve for books.*

The demand curve represents the WTP of buyers; similarly supply depends on the sellers' *willingness to accept* (WTA) money in return for books. The supply of second-hand books comes from students who have previously completed the course, who will differ in the amount they are willing to accept—that is, their reservation price. Recall from Unit 5 that Angela was willing to enter into a contract with Bruno only if it gave her as least as much utility as her reservation option (no work and survival rations); here the reservation price of a potential seller represents the value to her of keeping the book, and she will be willing to sell only for a price at least that high. Poorer students (who are keen to sell so that they can afford other books) and those no longer studying economics may have lower reservation prices. Again, online auctions like eBay allow sellers to specify their WTA.

We can draw a supply curve by lining up the sellers in order of their reservation prices (their WTAs): see Figure 8.2. We put the sellers who are most willing to sell—those who have the lowest reservation prices—first, so the graph of reservation prices slopes upward.

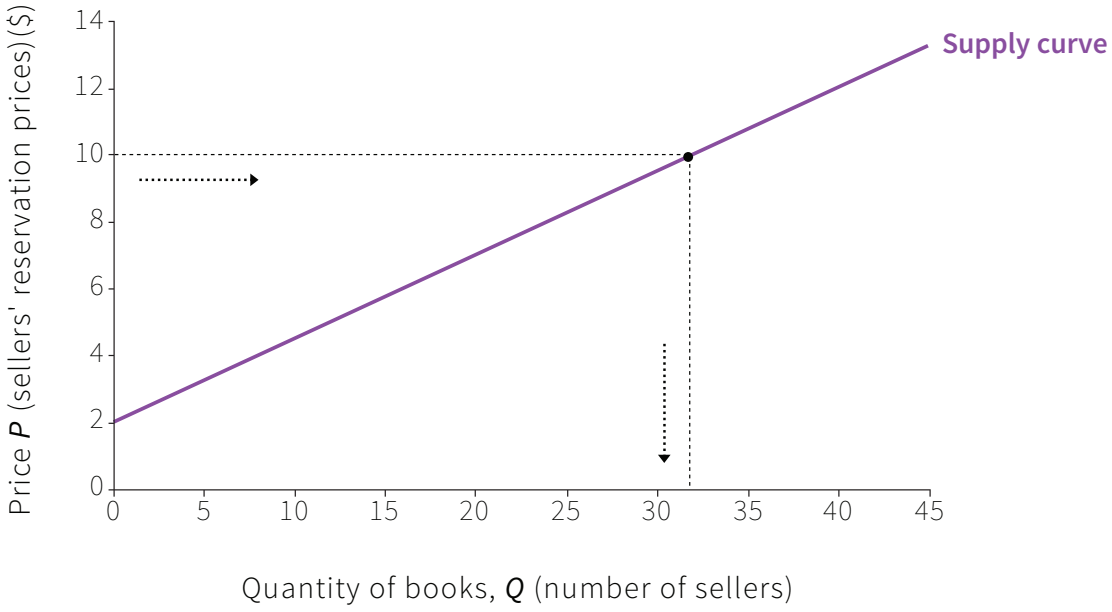


Figure 8.2 *Supply curve for books.*

The first seller has a reservation price of $2, and will sell at any price above that. The 20th seller will accept $7…… and the 40th seller's reservation price is $12. If you choose a particular price, say $10, the graph shows how many books would be supplied (*Q*) at that price: in this case, it is 32. The supply curve slopes upward: the higher the price, the more students will be willing to sell.

For any price, the supply curve shows the number of students willing to sell at that price—that is, the number of books that will be supplied to the market. Notice that we have drawn the supply and demand curves as straight lines for simplicity. In practice they are more likely to be curves, with shapes depending on how valuations of the book vary among the students.

DISCUSS 8.1: SELLING STRATEGIES AND RESERVATION PRICES

Imagine that you are planning to move to a city with good public transport and restricted parking, so you wish to sell your car. You are considering three possible methods:

- Advertise it in the local newspaper
- Take it to a car auction
- Offer it to a second-hand car dealer

1. Would your reservation price be the same in each case?
2. Why?
3. If you used the first method, would you advertise it at your reservation price?
4. Which method do you think would result in the highest sale price?
5. Which method would you choose?

8.2 THE MARKET AND THE EQUILIBRIUM PRICE

What would you expect to happen in the market for this textbook? That will depend on the market institutions that bring buyers and sellers together. If students have to rely on word-of-mouth, then when a buyer finds a seller they can try to negotiate a deal that suits both of them. But each buyer would like to be able to find a seller with a low reservation price, and each seller would like to find a buyer with high willingness to pay. Before concluding a deal with one trading partner they would like to know about other trading opportunities.

Traditional market institutions often brought many buyers and sellers together in one place. Many of the world's great cities grew up around marketplaces and bazaars along ancient trading routes such as the *Silk Road* between China and the Mediterranean. In the Grand Bazaar of Istanbul, one of the largest and oldest covered

markets in the world today, shops selling carpets, gold, leather and textiles cluster together in different areas. In medieval towns and cities it was common for makers and sellers of particular goods to set up shops close to each other, so customers knew where to find them. The City of London is now a financial centre, but evidence of trades once carried out there can be found in surviving street names: *Pudding Lane, Bread Street, Milk Street, Threadneedle Street, Ropemaker Street, Poultry*, and *Silk Street*.

With modern communications, sellers can advertise their goods and buyers can more easily find out what is available, and where. But in some cases it is still convenient for many buyers and sellers to meet together; large cities have markets for meat, fish, vegetables or flowers, where buyers can inspect and compare the quality of the produce. In the past, markets for second-hand goods often involved specialist dealers, but nowadays sellers can contact buyers directly through online marketplaces such as eBay. Returning to our example of second-hand textbooks, local online sites now help students sell books to others in their university.

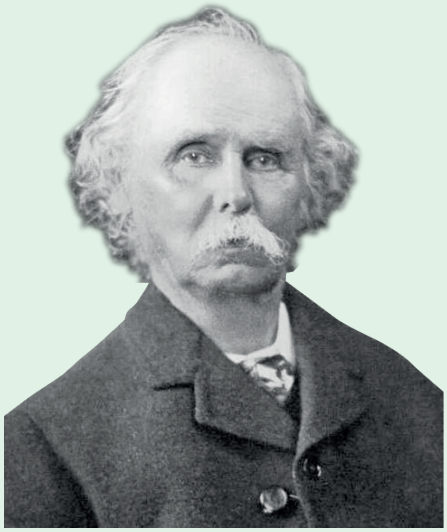
At the end of the 19th century, the economist Alfred Marshall introduced his model of supply and demand using an example that is quite similar to our case of second-hand books. Most English towns had a Corn Exchange (also known as a grain exchange)—a building where farmers met with merchants to sell their grain. Marshall described how the supply curve of grain would be determined by the prices that farmers would be willing to accept, and the demand curve by the willingness to pay of merchants. Then he argued that, although the price "may be tossed hither and thither like a shuttlecock" in the "higgling and bargaining" of the market, it would never be very far from the particular price at which the quantity demanded by merchants was equal to the quantity the farmers would supply.

Marshall called the price that equated supply and demand the *equilibrium price*. If the price was above the equilibrium, farmers would want to sell large quantities of grain. But few merchants would want to buy; there would be *excess supply*. Then, even the merchants who were willing to pay that much would realise that farmers would soon have to lower their prices and would wait until they did. Similarly if the price was below the equilibrium, sellers would prefer to wait rather than sell at that price. If, at the going price, the amount supplied did not equal the amount demanded, Marshall reasoned that some sellers or buyers could benefit by charging some other price. The prevailing price would not be what we now term a Nash equilibrium. So the price would tend to settle at the equilibrium level where demand and supply were equated.

Marshall's argument was based on the assumption that all the grain was of the same quality. His supply and demand model can be applied to markets in which all sellers are selling the identical goods, so buyers are equally willing to buy from any seller. If the farmers all had grain of different qualities, they would be more like the sellers of differentiated products in Unit 7, who could set their own prices.

GREAT ECONOMISTS

ALFRED MARSHALL



Alfred Marshall (1842-1924) was a founder—along with Léon Walras—of what is termed the *neoclassical school* of economics. His *Principles of Economics*, first published in 1890, was the standard introduction to economics textbook for English speaking students for 50 years. An excellent mathematician, Marshall provided new foundations for the analysis of supply and demand by formulating the workings of markets and firms using calculus to express such central concepts as *marginal costs* and *marginal utilities*. The concepts of consumer and producer surplus are due to Marshall. His conception of economics as an attempt to "understand the influences exerted on the quality and tone of a man's life by the manner in which he earns his livelihood…" is close to our own definition of the field.

But much of the wisdom in Marshall's text, sadly, has rarely been taught by his followers. Marshall paid attention to facts. His observation that large firms could produce at lower unit costs than small firms was integral to his thinking, but it never found a place in the neoclassical school. This may be because, if the average cost curve is downward-sloping until firms are very large, there will be a kind of winner-take-all competition in which a few large firms emerge as winners *with the power to set prices*, rather than taking the going price as a given. We return to this problem in Unit 10 and Unit 20.

Marshall would also have been distressed that homo economicus (we questioned his existence in Unit 4) became the main actor in textbooks written by the followers of the neoclassical school. He insisted that:

"Ethical forces are among those of which the economist has to take account. Attempts have indeed been made to construct an abstract science with regard to the actions of an economic man who is under no ethical influences and who pursues pecuniary gain… selfishly. But they have not been successful."
— Alfred Marshall, *Principles of Economics* (1890)

While advancing the use of mathematics in economics, he also cautioned against its misuse. In a letter to A. L. Bowley, a fellow mathematically inclined economist, he explained his own "rules" as follows:

1. Use mathematics as a shorthand language, rather than as an engine of inquiry
2. Keep to them till you have done
3. Translate into English
4. Then illustrate by examples that are important in real life
5. Burn the mathematics
6. If you can't succeed in 4, burn 3: "This I do often."

Marshall was Professor of Political Economy at the University of Cambridge between 1885 and 1908. In 1896 he circulated a pamphlet to the University Senate objecting to a proposal to allow women to be granted degrees. Marshall prevailed and women would wait until 1948 before being granted academic standing at Cambridge on a par with men.

But his work was motivated by a desire to improve the material conditions of working people. Readers of Principles had no doubt about what he thought was the main task of economics:

"Now at last we are setting ourselves seriously to inquire whether it is necessary that there should be any so called lower classes at all: that is whether there need be large numbers of people doomed from their birth to hard work in order to provide for others the requisites of a refined and cultured life, while they themselves are prevented by their poverty and toil from having any share or part in that life. ...[T]he answer depends in a great measure upon facts and inferences, which are within the province of economics; and this is it which gives to economic studies their chief and their highest interest."
— Alfred Marshall, *Principles of Economics* (1890)

Would Marshall would now be satisfied with the contribution that modern economics has made to creating a more just economy?

To apply this model to the market for the textbook, we assume that all the books are the same (although in practice some may be in better condition than others) and that a potential seller can advertise a book for sale by announcing its price on a local website. As at the Corn Exchange, we would expect that most trades would occur at similar prices. Buyers and sellers can easily observe all the advertised prices, so if some books were advertised at $10 and others at $5, buyers would be queuing up to pay $5, and these sellers would quickly realise that they could charge more, while no-one would want to pay $10 so these sellers would have to lower their price.

We can find the equilibrium price by drawing the supply and demand curves on one diagram, as in Figure 8.3. At a price $P^* = \$8$, the supply of books is equal to demand: 24 buyers are willing to pay $8, and 24 sellers are willing to sell. The equilibrium quantity is $Q^* = 24$.

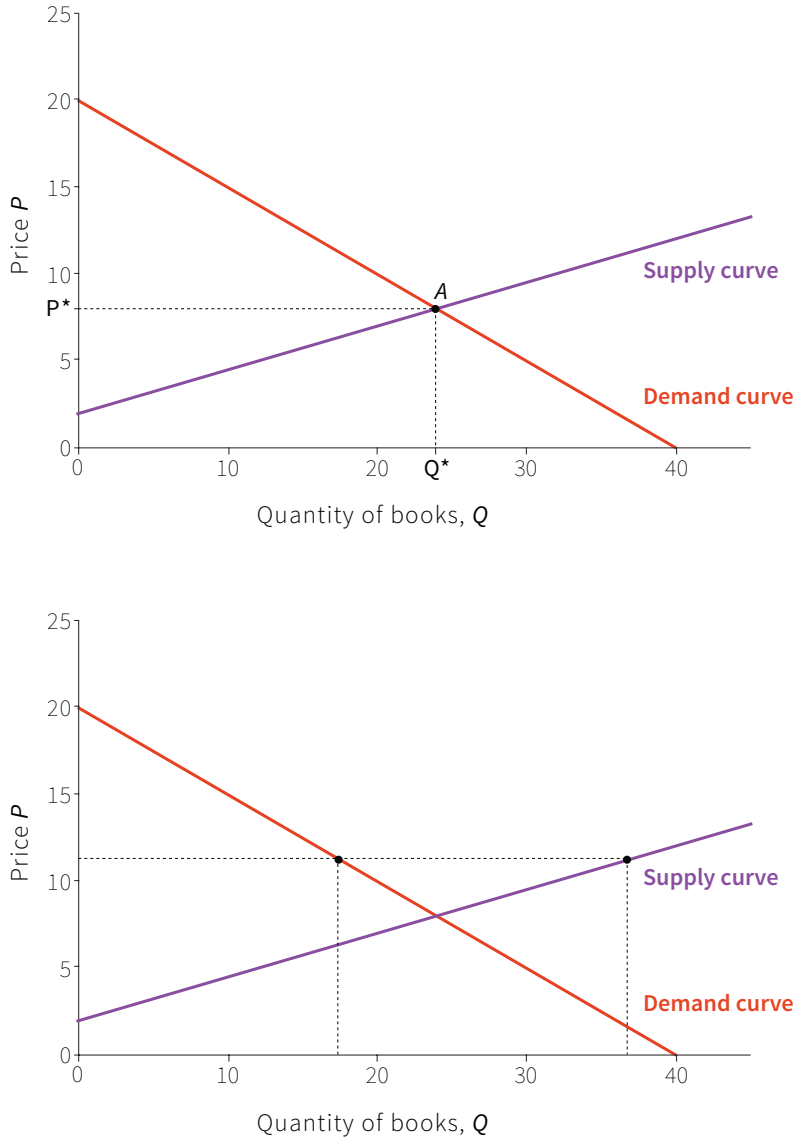


Figure 8.3 *Equilibrium in the market for second-hand books.*

The *market clears* at a price of $8; that is, supply is equal to demand at this price, so all buyers who want to buy and all sellers who want to sell can do so. The market is in equilibrium. In everyday language something is in equilibrium if the forces acting on it are in balance, so that it remains still. Remember Fisher's hydraulic model of price determination from Unit 2: changes in the economy caused water to flow through the apparatus until it reached an equilibrium with no further tendency for prices to change. We say that a market is in equilibrium if the actions of buyers and sellers have no tendency to change the price or the quantities bought and sold (as long as there is no change in market conditions such as the numbers of potential buyers and sellers, and how much they value the good). At the equilibrium price for the

textbook, all those who wish to buy or sell are able to do so, so there is no tendency for change (although when the conditions for equilibrium are not met, automatic price-setting algorithms might decide to sell a book for $23m).

Price-taking

Will the market always be in equilibrium? As we have seen, Marshall argued that the price would not deviate far from the equilibrium level. At a higher price, for example, there would be excess supply, and sellers would profit by lowering their prices. In this unit we study competitive market equilibria. In Unit 9 we will look at when and how prices change when the market is not in equilibrium.

In the market equilibrium that we have described for the textbook, individual students have to accept the prevailing price in the market, determined by the supply and demand curves. No one would trade with a student asking a higher price or offering a lower amount, because anyone they wanted to trade with would find an alternative buyer or seller instead. The participants in this market are *price-takers*, because there is sufficient competition from other buyers and sellers that the best they can do is to trade at the same price. Any buyer or seller is of course free to choose a different price, but they cannot benefit by doing so.

PRICE-TAKER

Buyers and sellers are *price-takers* if they cannot benefit by choosing a different price from the one at which everyone else is transacting.

We have seen other examples where market participants do not behave as price-takers: the producer of a differentiated product can set its own price because it has no close competitors. In Unit 6 we saw that, because labour contracts are incomplete, an employer may not seek to pay the lowest possible price for a worker, but instead choose to set a higher wage. Notice, however, that although the sellers of differentiated products are price-setters, the buyers in the models in Unit 7 were price-takers: since there are so many consumers wanting to buy breakfast cereals, an individual consumer has no power to negotiate a more advantageous deal, but simply has to accept the price that all other consumers are paying.

COMPETITIVE EQUILIBRIUM

A market is in *competitive equilibrium* if all buyers and sellers are price-takers, and at the prevailing market price, the quantity supplied is equal to the quantity demanded.

In this unit we will study market equilibria where both buyers and sellers are price-takers. We expect to see price-taking on both sides of the market where there are many sellers selling the identical goods, and many buyers wishing to purchase them. Sellers are forced to be price-takers by the presence of other sellers, and of buyers who always choose the seller with the lowest price. If a seller tried to set a higher price, buyers would simply go elsewhere.

Similarly buyers are price-takers when there are plenty of other buyers, and sellers willing to sell to whoever will pay the highest price. On both sides of the market, competition eliminates bargaining power. We will describe the equilibrium in such a market as a *competitive equilibrium*.

A competitive market equilibrium is a Nash equilibrium because given what all other actors are doing (trading at the equilibrium price) no actor can do better than to continue what he or she is doing (also trading at the equilibrium price).

DISCUSS 8.2: PRICE-TAKERS

Think about some of the goods you buy: perhaps different kinds of food, clothes, transport tickets, electronic goods.

1. Are there many sellers of these goods?
2. Do you try to find the one with the lowest price?
3. If not, why not?
4. For which goods would price be your main criterion?
5. Use your answers to help you decide whether the sellers of these goods are price-takers. Are there are goods for which you, as a buyer, are not a price-taker?

8.3 PRICE-TAKING FIRMS

In the second-hand textbook example, both buyers and sellers are individual consumers. Now we look at markets where the sellers are firms. We know from Unit 7 how firms choose their price and quantity when producing differentiated goods; and we saw that if other firms made similar products, their choice of price would be restricted (the demand curve for their own product would be almost flat) because raising the price would cause consumers to switch to other, similar, brands.

If there were many firms producing identical products, and consumers could easily switch from one firm to another, then firms will be price-takers in equilibrium. They will be unable to benefit from attempting to trade at a price different from the prevailing price.

To see how price-taking firms behave, consider a city where many small bakeries produce bread and sell it direct to consumers. Figure 8.4 shows what the market demand curve—the total daily demand for bread of all consumers in the city—might look like. It is downward-sloping as usual because, at higher prices, fewer consumers will be willing to pay.

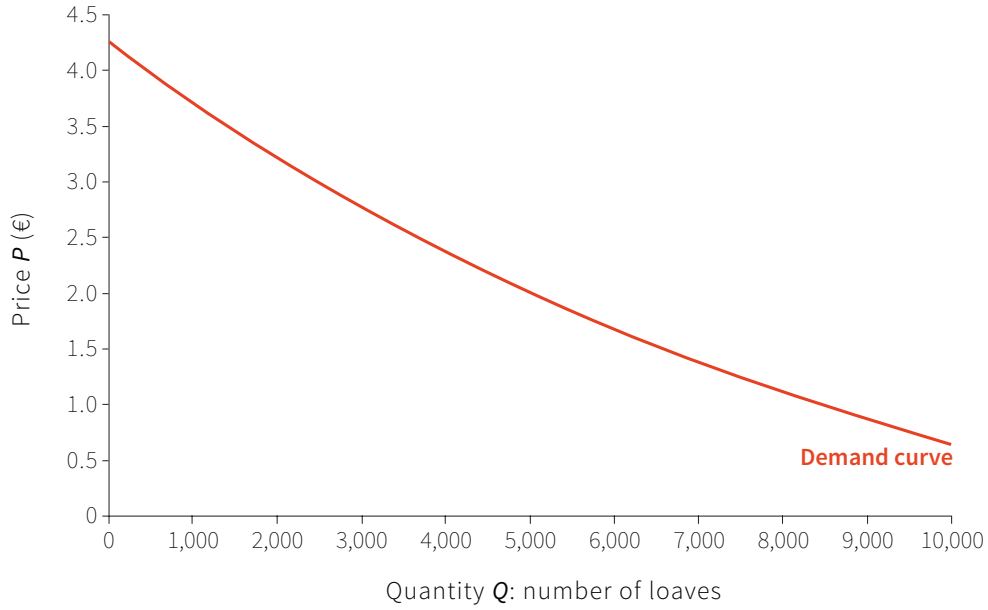


Figure 8.4 *The market demand curve for bread.*

Suppose that you are the owner of one small bakery. You have to decide what price to charge and how many loaves to produce each morning. Suppose that neighbouring bakeries are selling loaves identical to yours at €2.35. This is the prevailing market price, and you will not be able to sell loaves at a higher price than other bakeries, because no-one would buy—you are a price-taker.

Your marginal costs increase with your output of bread. When the quantity is small the marginal cost is low, close to €1; having installed mixers, ovens and other equipment, and employed a baker, the additional cost to produce a loaf of bread is relatively small; but the average cost of a loaf is high. As the number of loaves per day increases the average cost falls, but marginal costs begin to rise gradually because you have to employ extra staff and use equipment more intensively. At higher quantities the marginal cost is above the average cost; then average costs rise again.

The marginal and average cost curves are drawn in Figure 8.5. As in Unit 7, costs include the opportunity cost of capital. If price were equal to average cost ($P = AC$) your economic profit would be zero. You, the owner, would obtain a normal return on your capital. So the average cost curve (in Figure 8.5, it is furthest to the left) is the zero-economic-profit curve. The isoprofit curves show combinations of price and quantity at which you would receive higher levels of profit. As we explained in Unit 7, isoprofit curves slope down where price is above marginal cost, and up where price is below marginal cost, so the marginal cost curve passes through the lowest

point on each isoprofit curve. If price is above marginal cost, total profits can remain unchanged only if a larger quantity is sold for a lower price. Similarly, if price is below marginal cost, total profits can remain unchanged only if a larger quantity is sold for a higher price.

Figure 8.5 demonstrates how to make your decision. Like the firms in Unit 7 you face a constrained optimisation problem. You want to find the point of maximum profit in your feasible set.

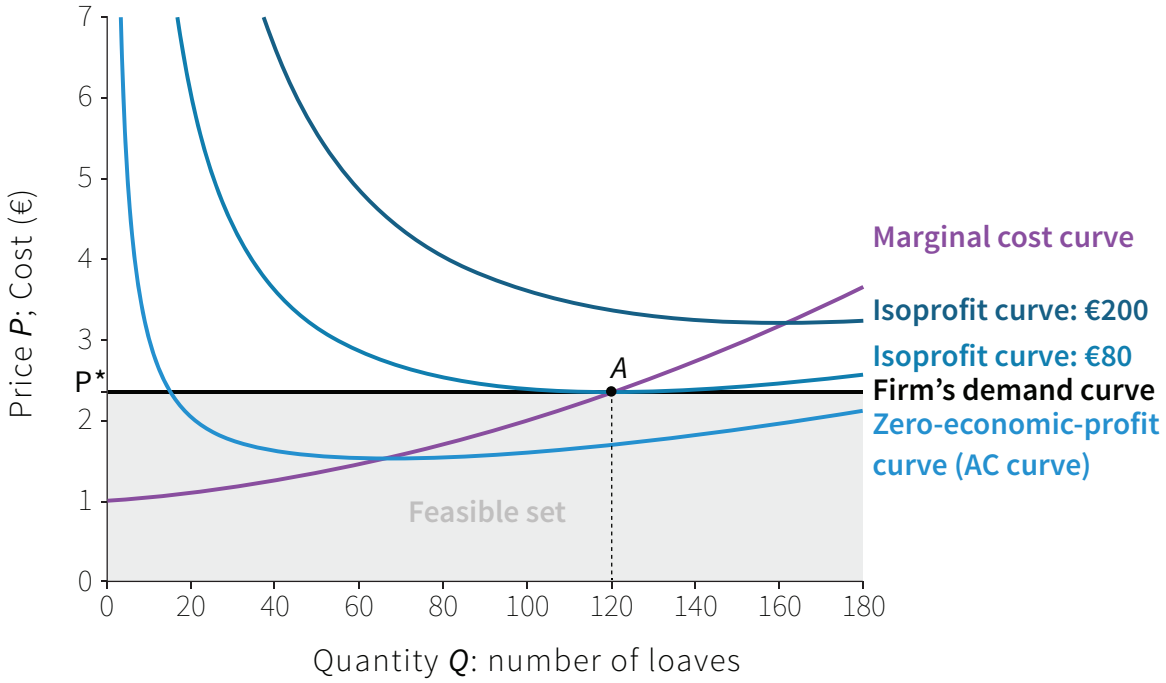


Figure 8.5 *The profit-maximising price and quantity for the bakery.*

Because you are a price-taker, the feasible set is all points where price is less than or equal to €2.35, the market price. Your optimal choice is $P^* = €2.35$, $Q^* = 120$, where the isoprofit curve is tangent to the feasible set. The problem looks similar to the one for *Beautiful Cars* in Unit 7 except that, for a price-taker, the demand curve is completely flat. For your bakery, it is not the market demand curve in Figure 8.4 that affects your own demand; it is the price charged by your competitors. This is why the horizontal line at P^* in Figure 8.5 is labelled as the firm's demand curve. If you charge more than P^* your demand will be zero; at P^* or less you can sell as many loaves as you like.

Figure 8.5 illustrates a very important characteristic of price-taking firms. They choose to produce a quantity at which the marginal cost is equal to the market price ($MC = P^*$). This is always true. For a price-taking firm, the demand curve for its own output is a horizontal line at the market price; so maximum profit is achieved at a point on the demand curve where the isoprofit curve is horizontal. And we know from Unit 7, that where isoprofit curves are horizontal, the price is equal to the marginal cost.

Another way to understand why a price-taking firm produces at the level of output where $MC = P^*$ is to think about what would happen to its profits if it deviated from this point. If the firm were to increase output to a level where $MC > P^*$, the last unit would cost more than P^* to make, so the firm would lose on this unit and would make higher profits by reducing output. If it were to produce where $MC < P^*$, it could produce at least one more unit and sell it at a profit. Therefore it should raise output as far as the point where $MC = P^*$. This is where profits are maximised.

PRICE-TAKING FIRM

A *price-taking firm* maximises profit by choosing a quantity where the marginal cost is equal to the market price ($MC = P^*$) and selling at the market price P^*.

This is an important result, which you should remember, but you need to be careful with it. When we make statements like "for a price-taking firm, price equals marginal cost", we do *not* mean that the firm chooses a price equal to its marginal cost. We mean the opposite: it is a price-taker, so it accepts the market price, and *chooses its quantity* so that the marginal cost is equal to the price.

Put yourself in the position of the bakery owner again. What would you do if the market price changed? Figure 8.6 demonstrates that as prices change you would choose different points on the marginal cost curve.

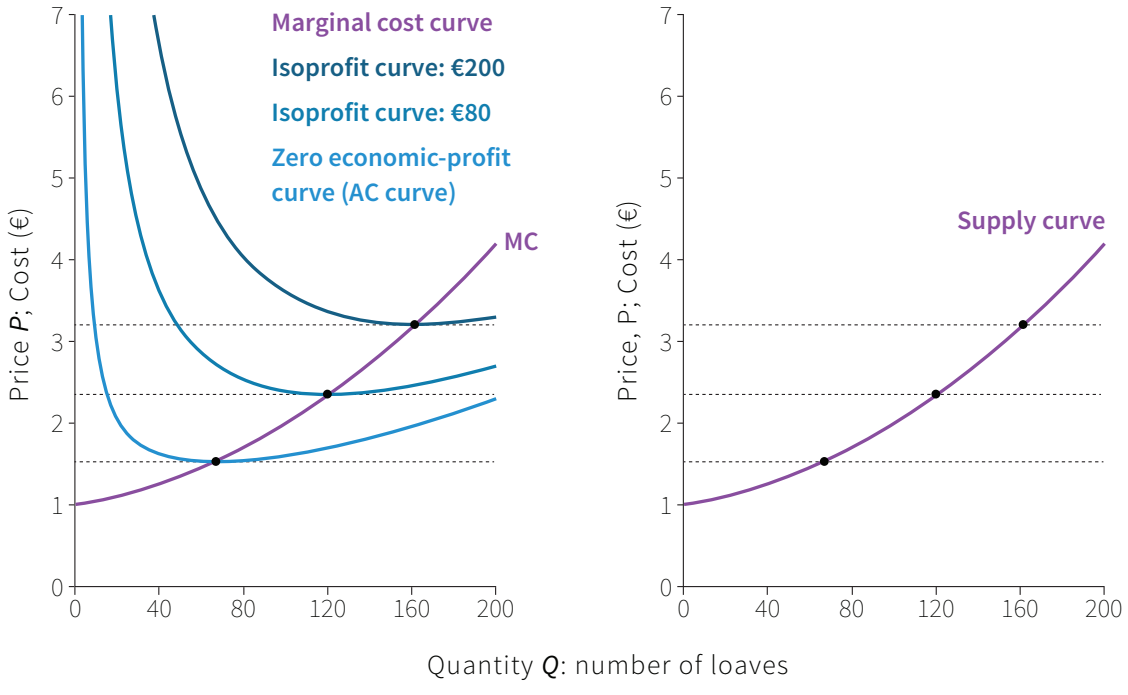


Figure 8.6 *The firm's supply curve.*

When the market price is €2.35, you supply 120 loaves. What would you do if the price changed? If P^* were to rise to €3.20, you could reach a higher isoprofit curve. To maximise profit you should produce 163 loaves per day. If the price falls to €1.52 you could reach only the lightest blue curve. Your best choice would be 66 loaves, and your economic profit would be zero. In each case, you choose the point on your marginal cost curve where *MC = market price*. Your marginal cost curve is your supply curve.

For a price-taking firm, *the marginal cost curve is the supply curve*: for each price it shows the profit-maximising quantity: that is the quantity that the firm will choose to supply.

Notice, however, that if the price fell below €1.52 you would be making a loss. The supply curve shows how many loaves you should produce to maximise profit, but when the price is this low, the economic profit is nevertheless negative. On the supply curve, you would be minimising your loss. If this happened you would have to decide whether it was worth continuing to produce bread. Your decision depends on what you expect to happen in the future:

- If you expect market conditions to remain bad, it might be best to sell up and leave the market—elsewhere, you could obtain a better return on your capital.
- If you expect the price to rise soon, you might be willing to incur some short-term losses—and it might be worth continuing to produce bread if the revenue helped you to cover the costs of maintaining your premises and retaining staff.

8.4 MARKET SUPPLY AND EQUILIBRIUM

The market for bread in the city has many consumers and many bakeries. Let's suppose there are 50 bakeries. Each one has a supply curve corresponding to its own marginal cost curve, so we know how much it will supply at any given market price. To find the market supply curve, we just add up the total amount that all the bakeries, together, will supply at each price.

Figure 8.7 shows how this works if all the bakeries have the same cost functions. We work out how much one bakery would supply at a given price, then multiply by 50 to find total market supply at that price.

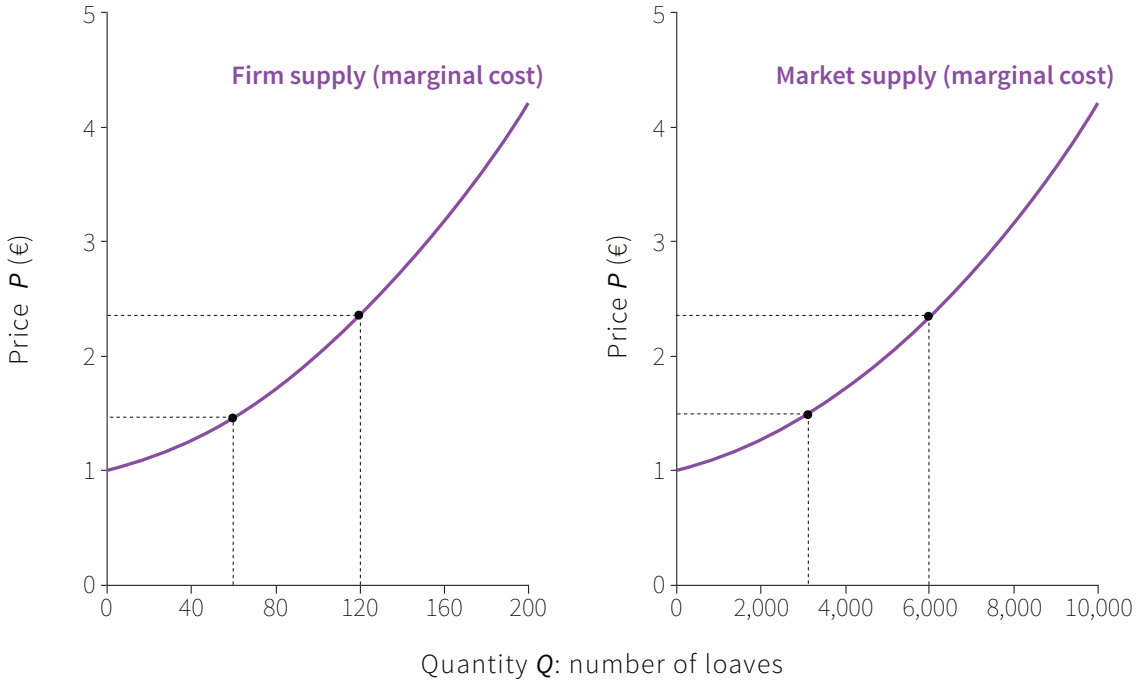


Figure 8.7 *The firm and market supply curves.*

There are 50 bakeries, all with the same cost functions. If the market price is €2.35, each bakery will produce 120 loaves. When P= €2.35 each bakery supplies 50 loaves, and the market supply is *50 x 120 = 6,000* loaves. At a price of €1.52 they each supply 66 loaves, and market supply is 3,300. The market supply curve looks like the firm's supply curve, except that the scale on the horizontal axis is different. If the bakeries had different cost functions, then at a price of €2.35 some bakeries would produce more loaves than others, but we could still add them together to find market supply.

The market supply curve shows the total quantity that all the bakeries together would produce at any given price. Also, it represents the marginal cost of producing a loaf, just as the firm's supply curve does. For example, if the market price is €2.75, total market supply is 7,000. For every bakery, the marginal cost—the cost of producing one more loaf—is €2.75. And that means that the cost of producing the 7,001st loaf in the market is €2.75, whichever firm produces it. So *the market supply curve is the market's marginal cost curve*. If your course includes calculus, our Leibniz supplement explains how the market supply curve is derived from the individual firms' supply curves.

Now we know both the demand curve (Figure 8.4), and the supply curve (Figure 8.7) for the bread market as a whole. Figure 8.8 shows that the equilibrium price is exactly €2.00. At this price, we say the market *clears*: consumers demand 5,000 loaves per day, and firms supply 5,000 loaves per day.

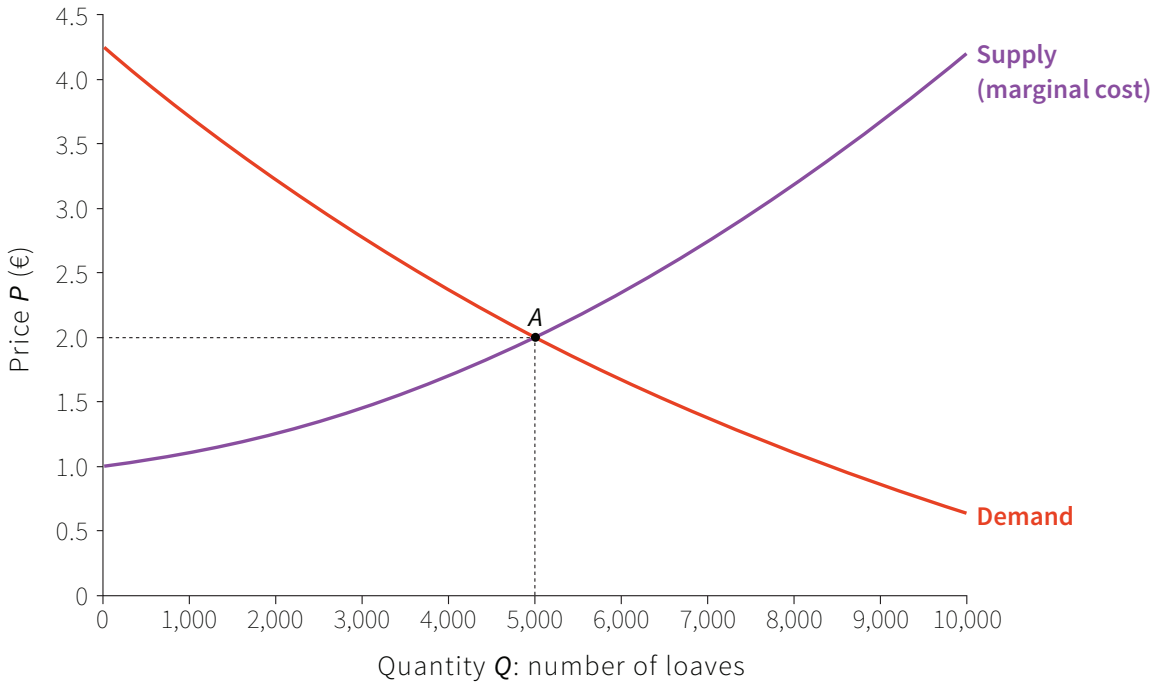


Figure 8.8 *Equilibrium in the market for bread.*

In the market equilibrium, each bakery is producing on its marginal cost curve, at the point where its marginal cost is €2.00. To find out how to calculate the market price algebraically, see this Leibniz supplement. If you look back to the isoprofit curves in Figure 8.6, you will see that the firm is above its average cost curve—the isoprofit curve where economic profits are zero. So the owners of the bakeries are receiving economic rents—profit in excess of normal profit. Whenever there are economic rents, there is an opportunity for market participants to benefit by taking an action. In this case, we might expect the economic rents to attract other bakeries into the market. We shall see presently how the entry of more firms would increase the supply of bread in the longer term, and could eventually reduce economic profits to zero, eliminating rents.

8.5 COMPETITIVE EQUILIBRIUM: GAINS FROM TRADE, ALLOCATION AND DISTRIBUTION

Buyers and sellers of bread voluntarily engage in trade because both benefit. Their mutual benefits from the equilibrium allocation can be measured by the consumer and producer surpluses introduced in Unit 7. Any buyer whose willingness to pay for a good is higher than the market price receives a surplus: the difference between the WTP and the price paid. Similarly, if the marginal cost of producing a good is below

the market price, the producer receives a surplus. Figure 8.9a shows how to calculate the total surplus (the gains from trade) at the competitive equilibrium in the market for bread, in the same way as we did for the markets in Unit 7.

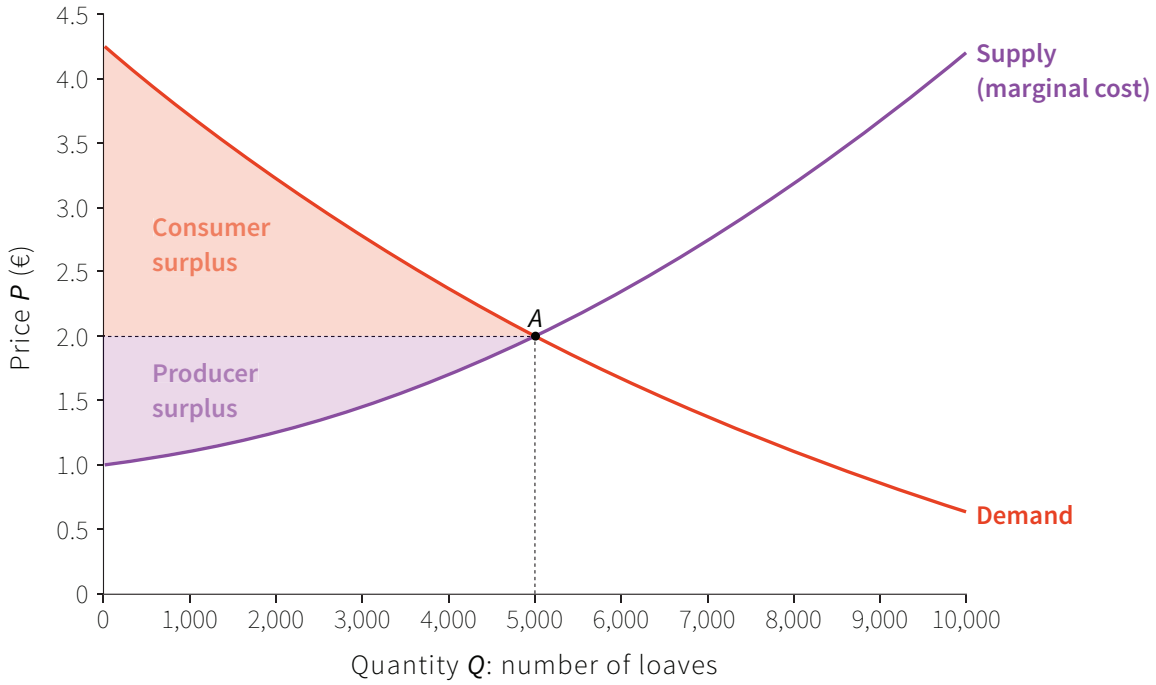


Figure 8.9a *Equilibrium in the bread market: gains from trade.*

At the equilibrium price of €2 in the bread market, a consumer who is willing to pay €3.50 obtains a surplus of €1.50. The shaded area above €2 shows *total consumer surplus*—the sum of all the buyers' gains from trade. Remember from Unit 7 that the producer's surplus on a unit of output is the difference between the price at which it is sold, and the marginal cost of producing it. The marginal cost of the 2,000th loaf is €1.25; since it is sold for €2, the producer obtains a surplus or €0.75.The shaded area below €2 is the sum of the bakeries' surpluses on every loaf that they produce. The whole shaded area shows the sum of all gains from trade in this market, known as the *total surplus*.

Figure 8.9a shows that, when the market for bread is in equilibrium with the quantity of loaves supplied equal to the quantity demanded, the total surplus is the whole of the area below the demand curve and above the supply curve. The competitive equilibrium allocation of bread maximises the total surplus. However, if fewer than 5,000 loaves were produced in the bread market, there would be consumers without bread who would be willing to pay more than the cost of producing another loaf, so there would be unexploited gains from trade. Figure 8.9b shows what the total surplus would be if the bakeries produced only 4,000 loaves, and sold them at €2.00 each. The gains from trade in the market would be lower; there would be a *deadweight*

loss equal to the triangle-shaped area. Producers would be missing out on potential profits, and some consumers would be unable to obtain the bread they were willing to pay for.

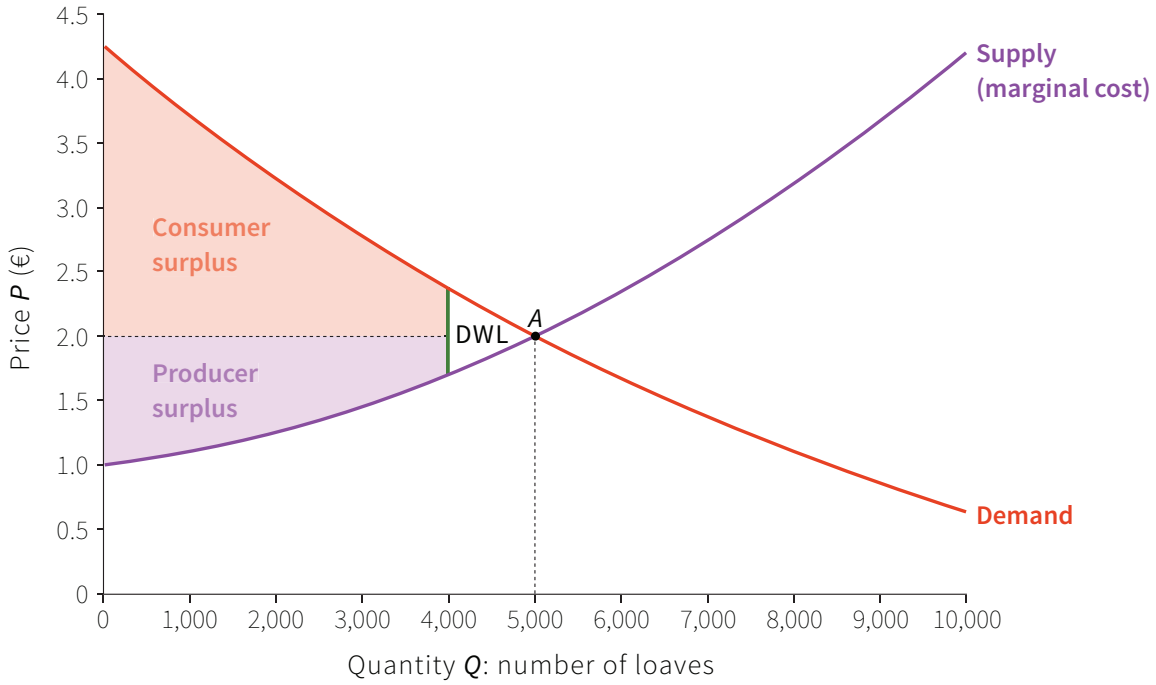


Figure 8.9b *Deadweight loss.*

And there are no gains to be made from more than 5,000 loaves, because none of the other consumers is willing to pay more than they would cost to make. At the equilibrium, all of the potential gains from trade are exploited.

This property—that the combined consumer and producer surplus is maximised at the point where supply equals demand—holds in general: if both buyers and sellers are price-takers, the equilibrium allocation maximises the sum of the gains achieved by trading in the market, relative to the original allocation. We demonstrate this result in our Einstein on total surplus and WTP. The concept of deadweight loss applies whenever there are gains from exchange: even at Christmas.

If either buyers or sellers are *not* price-takers, there may be a deadweight loss. For example, in Unit 7 the supplier set the price of a differentiated good. The price a supplier chose was above the marginal cost of the good, and the quantity was too low; we saw that this caused a deadweight loss.

The producer of a differentiated good has bargaining power (market power) because no-one else produces the same good. The firm uses its power to keep the price high, raising its own share of the surplus but lowering total surplus. In a competitive equilibrium, no individual has any bargaining power: when a particular buyer trades with a particular seller, each of them knows that the other can find an alternative trading partner willing to trade at the market price. Competition on both sides of the

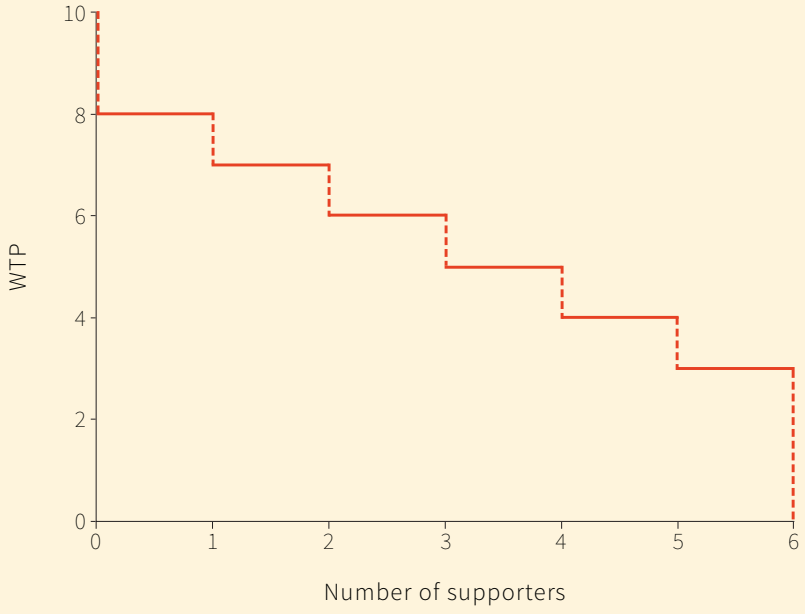
market eliminates the market power of both buyers and sellers. Sellers can't raise the price because of competition from other sellers; competition from other buyers prevents buyers from lowering it.

In the scenario in Unit 5, both Angela and Bruno had bargaining power, since neither of them faced any competition. But competition from other workers wanting to work for Bruno would eliminate Angela's bargaining power. Then Bruno could make a take-it-or-leave-it offer, obtaining the whole of the surplus, while Angela, with no power to refuse, would obtain only her reservation utility.

At the competitive equilibrium allocation in the bread market, it is not possible to make any of the consumers or firms better off (that is, to increase the surplus of any individual) without making at least one of them worse off. Provided that what happens in this market does not affect anyone other than the participating buyers and sellers, we can say the equilibrium allocation is Pareto efficient (we will consider below a case where this is not true, when noise from the bakeries disturbs their neighbours). To find out how to calculate the total surplus of a Pareto-efficient allocation using calculus, see this Leibniz supplement.

DISCUSS 8.3: MAXIMISING THE SURPLUS

Consider a market for the tickets to a football match. Six supporters of the Blue team would like to buy tickets; their valuations of a ticket (their WTP) are 8, 7, 6, 5, 4 and 3. The diagram below shows the demand "curve". Six supporters of the Red team already have tickets, for which their reservation prices (WTA) are 2, 3, 4, 5, 6 and 7.



1. Draw the supply and demand "curves" on a single diagram. (The supply curve is also a step function, like the demand curve.)
2. Show that, in equilibrium, four trades take place.
3. What is the equilibrium price?
4. Calculate the consumer (buyer) surplus by adding up the surpluses of the four buyers who trade.
5. Similarly calculate the producer (or seller) surplus.
6. Hence find the total surplus in equilibrium.
7. Suppose that the market operates through bargaining between individual buyers and sellers. Find a way of matching the buyers and sellers so that more than four trades occur. (Hint: suppose the highest WTP buyer buys from the highest WTA seller.)
8. In this case, work out the surplus from each trade.
9. How does the total surplus in this case compare with the equilibrium surplus?
10. Starting from the allocation of tickets you obtained through bargaining, in which at least five tickets are owned by Blue supporters, is there a way through further trade to make one of the supporters better off without making anyone worse off? (You can assume that none of them are football hooligans.)

Pareto efficiency comes about in our model of the bread market because of particular conditions that we have assumed:

- *The participants are price-takers*: They have no market power. For both buyers and sellers, competition from other buyers and sellers eliminates their power to affect the price at which they trade. Hence the suppliers will choose their output so that the marginal cost (the cost of the last unit produced) is equal to the market price. In contrast, a firm producing a differentiated product (such as a car) faces less competition and has the power to set its own price as a result. It chooses a higher price, above marginal cost.
- *The exchange of a loaf of bread for money is governed by a complete contract between buyer and seller*: If you find there is no loaf of bread in the bag marked "bread" when you get home, you can get your money back. Compare this with the incomplete employment contract in Unit 6, in which the firm can buy the worker's time, but cannot be sure how much effort the worker will put in. The firm chooses to pay more than the worker's reservation wage as an incentive for effort. In the labour market equilibrium there will be some unemployed workers who are willing to work at the equilibrium wage, so the allocation of jobs will not be Pareto efficient. The key difference between the bread market and the labour market is that the labour contract is incomplete: it can specify hours of work but not effort.

- *We assume that what happens in this market affects no one except the buyers and sellers*: On this basis, we claim that the allocation of bread is Pareto efficient. But if, for example, the early morning activities of bakeries disrupt the sleep of local residents, then there are additional costs of bread production; we ought to take the costs to the bakeries' neighbours into account when we assess Pareto efficiency. Then, we may conclude that the equilibrium allocation is not Pareto efficient after all. We will investigate this type of problem in Unit 10.

Even if we think that the market allocation is Pareto efficient, we should not conclude that it is necessarily a desirable one. Remember from Unit 5 that there are two criteria for assessing an allocation: efficiency and fairness. What can we say about fairness in the case of the bread market? We could examine the distribution of the gains from trade between producers and consumers: Figure 8.9 showed that both consumers and firms obtain a surplus, and in this example consumer surplus is slightly higher than producer surplus. You can see that this happens because the demand curve is relatively steep compared with the supply curve. Recall also from Unit 7 that a steep demand curve corresponds to a low elasticity of demand. Similarly the slope of the supply curve corresponds to the elasticity of supply: in Figure 8.9, demand is less elastic than supply.

In general:

- *The distribution of the total surplus between consumers and producers depends on the elasticities of demand and supply.*

We might also want to take into account the standard of living of participants in the market. For example, if a poor student buys a book from a rich student, we might think that an outcome in which the buyer paid less than the market price (closer to the seller's reservation price) would be better, because it would be fairer. Or, if the consumers in the bread market were exceptionally poor, we might decide that it would be better to pass a law setting a maximum bread price lower than €2.00 to achieve a fairer, although Pareto inefficient, outcome. In Unit 9 we will look at the effect of regulating markets in this way.

A competitive equilibrium allocation is Pareto efficient, which is often interpreted as a powerful argument in favour of markets as a mechanism for allocating resources. But we need to be careful not to overestimate the value of this result:

- *The allocation may not be Pareto efficient* if we have taken everything into account.
- *There are other important considerations* such as fairness.
- *Price-takers are hard to find in real life*: It is not as easy as you might think to find behaviour consistent with our simple model of the bread market (as we will see in section 8.9).

DISCUSS 8.4: SURPLUS AND DEADWEIGHT LOSS

1. Sketch a diagram to illustrate the competitive market for bread, showing the equilibrium where 5,000 loaves are sold at a price of €2.00.
2. Suppose that the bakeries get together to form a cartel. They agree to raise the price to €2.70, and jointly cut production to supply the number of loaves consumers demand at that price. Shade the areas on your diagram to show the consumer surplus, the producer surplus, and the deadweight loss caused by the cartel.
3. For what kinds of goods would you expect the supply curve to be highly elastic?
4. Draw diagrams to illustrate how the share of the gains from trade obtained by producers depends on the elasticity of the supply curve.

8.6 CHANGES IN SUPPLY AND DEMAND

Quinoa is a cereal crop grown on the *altiplano*, a high, barren plateau in the Andes of South America. It is a traditional staple food in Peru and Bolivia. In recent years, as its nutritional properties have become known, there has been a huge increase in demand from richer health-conscious consumers in Europe and North America. Figure 8.10 shows how the market has changed. You can see in Figures 8.10a and 8.10b that between 2001 and 2011 the price trebled, and production almost doubled. Figure 8.10c indicates the strength of the increase in demand: spending on imports of quinoa rose from just $2.4m to $43.7m in 10 years.

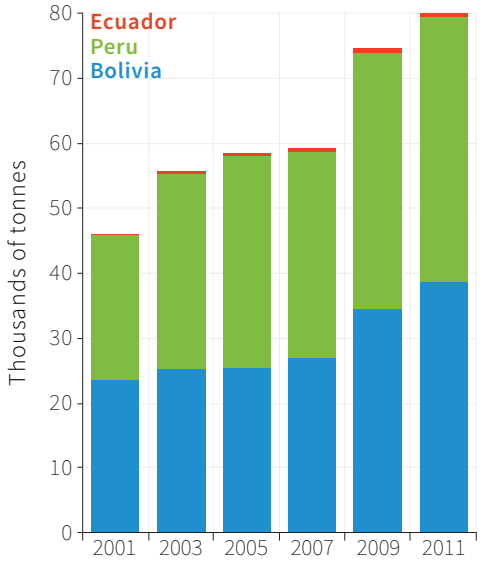


Figure 8.10a *The production of quinoa.*

Source: *Reyes, Jose Daniel, and Julia Oliver. 2013. 'Quinoa: The Little Cereal That Could.' The Trade Post. World Bank. November 22. Underlying data from Food and Agriculture Organization of the United Nations. 2015. 'FAOSTAT Database.' Accessed July.*

For the producer countries these changes are a mixed blessing: while their staple food has become expensive for poor consumers, farmers—who are amongst the poorest—are benefiting from the boom in export sales. Other countries are now investigating whether quinoa can be grown in different climates, and France and the US have become substantial producers.

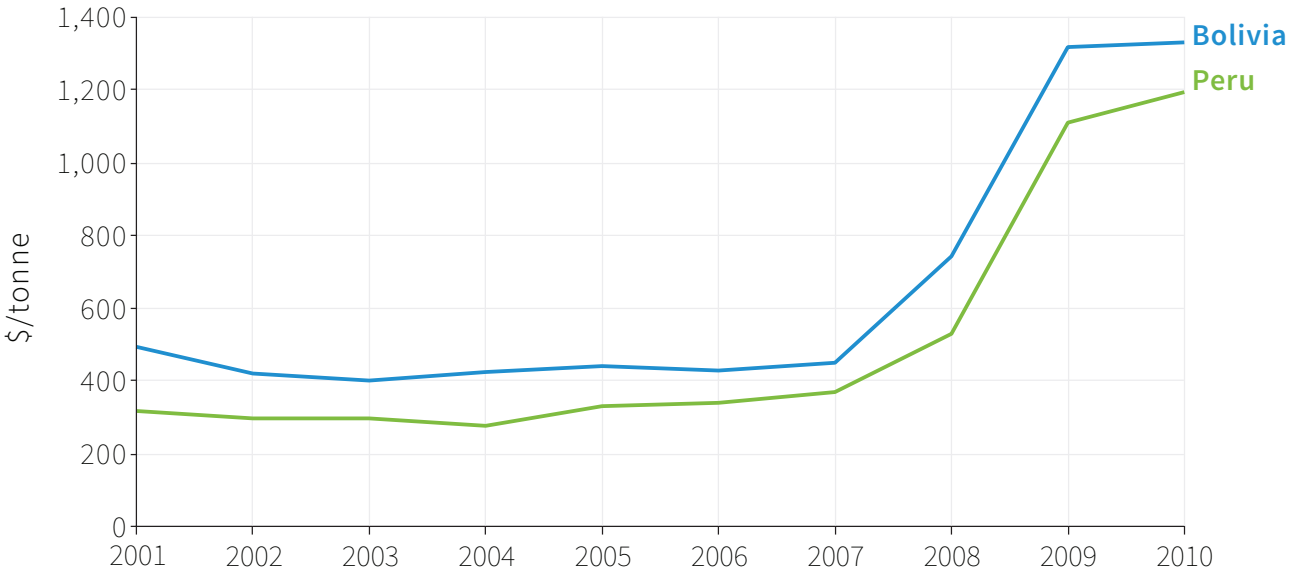


Figure 8.10b *Quinoa producer prices.*

Source: *Reyes, Jose Daniel, and Julia Oliver. 2013. 'Quinoa: The Little Cereal That Could.' The Trade Post. World Bank. November 22. Underlying data from Food and Agriculture Organization of the United Nations. 2015. 'FAOSTAT Database.' Accessed July.*

How can we explain the rapid increase in the price of quinoa? In this section and the next, we look at the effects of changes in demand and supply in our simple examples of books and bread, before returning at the end of section 8.7 to apply the analysis to the real-world case of quinoa.

In the market for the second-hand textbook, demand comes from new students enrolling on the course, and supply from students in the previous year. In Figure 8.11 we have plotted supply and demand for the book while the number of students enrolling

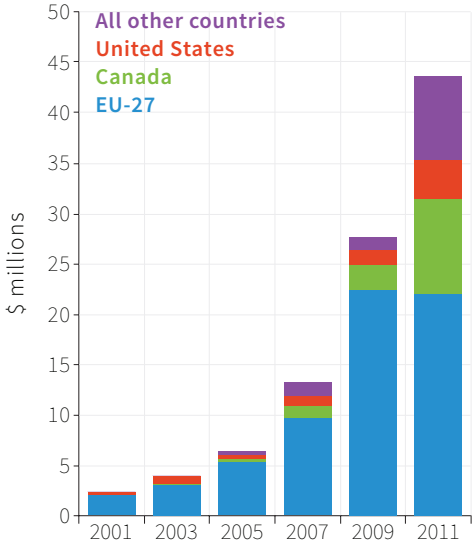


Figure 8.10c *Global import demand for quinoa.*

Source: *Reyes, Jose Daniel, and Julia Oliver. 2013. 'Quinoa: The Little Cereal That Could.' The Trade Post. World Bank. November 22. Underlying data from Food and Agriculture Organization of the United Nations. 2015. 'FAOSTAT Database.' Accessed July.*

remains stable at 40 per year. The equilibrium price is $8, and 24 books are sold, as shown by point A. Suppose that in one year the course became more popular. Figure 8.11 shows what would happen.

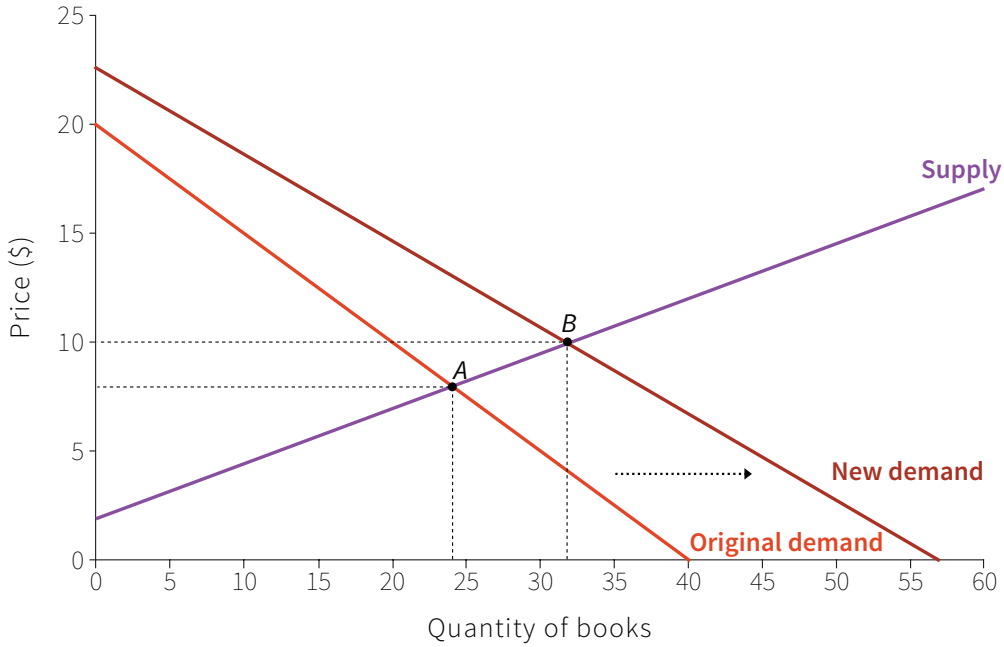


Figure 8.11 *An increase in demand for books.*

The increase in demand leads to a new equilibrium, in which 32 books are sold, for $10 each. At the original price, there would be excess demand and sellers would want to raise their prices. At the new equilibrium, both price and quantity are higher. Some students who would not have sold their books at $8 will now sell at a higher price. Notice, however, that although demand has increased, not all the students who would have bought at $8 purchase the book at the new equilibrium: those with WTP between $8 and $10 no longer want to buy.

When we say "increase in demand" it's important to be careful about exactly what has happened in this case:

- Demand is *higher at each possible price*, so the demand curve has shifted.
- In response to the shift there is a change in the price.
- This leads to an increase in the quantity supplied.
- This change is a movement along the supply curve.
- But the supply curve *has not shifted* (the number of sellers and their reserve prices have not changed), so we do not call this "an increase in supply".

In contrast, as an example of an increase in supply, think again about the market for bread in one city. Remember that the supply curve represents the marginal cost of producing bread. Suppose that bakeries discover a new technique that allows each

worker to make bread more quickly. This will lead to a fall in the marginal cost of a loaf at each level of output. In other words, the marginal cost curve of each bakery shifts down.

Figure 8.12 shows the original supply and demand curves for the bakeries. When the MC curve of each bakery shifts down, so does the market supply curve for bread. Look at Figure 8.12 to see what happens next.

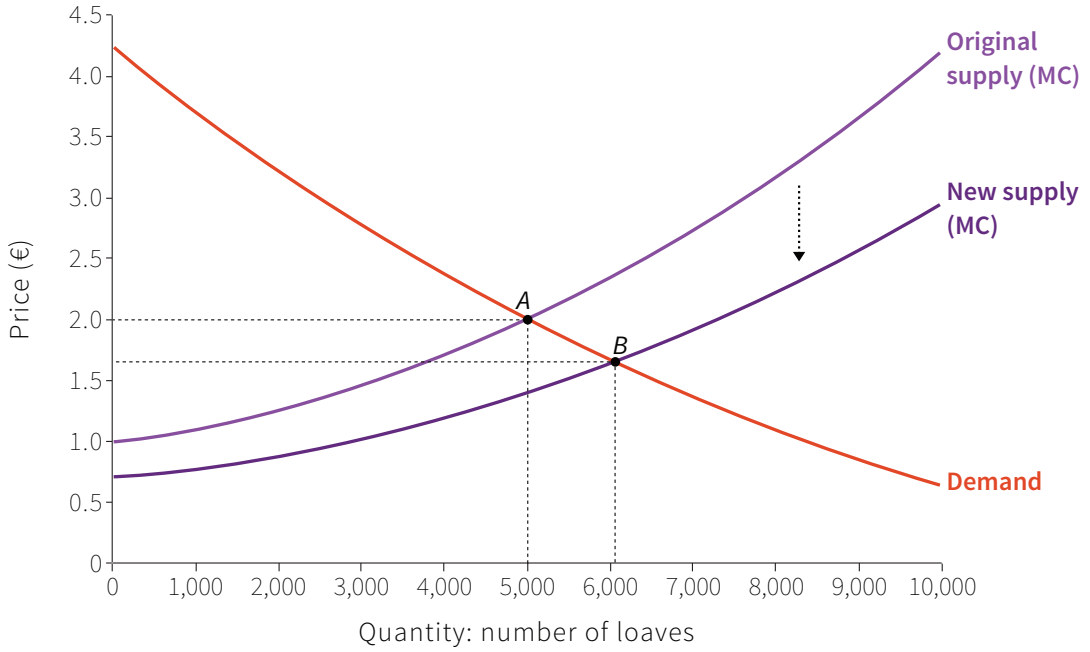


Figure 8.12 *An increase in the supply of bread (fall in MC).*

The city's bakeries start out at point *A*, producing 5,000 loaves and selling them for €2 each. The market supply curve then shifts because of the fall in the bakeries' marginal costs. The supply curve shifts down, because at each level of output, the marginal cost and therefore the price at which they are willing to supply bread is lower. The supply curve has shifted down. But another way to think of this change in supply is to say that the supply curve has shifted to the right. Since costs have fallen, the amount that bakeries will supply at each price is greater—an increase in supply. The effect of the fall in marginal cost is an increase in market supply; at the original price there is more bread than buyers want (excess supply). The bakeries would want to lower their prices. The new equilibrium market is at point *B* where more bread is sold, and the price is lower. The demand curve has not shifted, but the fall in price has led to an increase in the quantity of bread demanded, along the demand curve.

The improvement in the technology of breadmaking leads to:

- An increase in supply
- A fall in the price of bread
- A rise in the quantity sold

As in the example of an increase in demand, an adjustment of prices is needed to bring the market into equilibrium. Such shifts in supply and demand are often referred to as *shocks* in economic analysis. We start by specifying an economic model and find the equilibrium. Then we look at how the equilibrium changes when something changes—the model receives a shock. The shock is called *exogenous* because our model doesn't explain why it happened: the model shows the consequences, not the causes. This Leibniz supplement shows you how to model shifts in supply and demand mathematically.

Would you expect the market to adjust quickly to equilibrium? In this example it seems quite plausible, since the supply and demand curves are unlikely to change much from day to day. Bakeries that were left with unsold loaves would quickly adjust their prices and quantities to bring supply in line with demand. We will look at this process in more detail in Unit 9.

DISCUSS 8.5: BREAD, PRICES, SHOCKS AND REVOLUTION

Historians have usually attributed the wave of revolutions across Europe in 1848 to long-term socioeconomic factors and a surge of radical ideas. But a poor wheat harvest in 1845 lead to food shortages and sharp price rises—price shocks—in many European countries over the next three years. Helge Berger and Mark Spoerer, two economic historians, have investigated whether these short-term economic factors contributed to the sudden social and political changes that took place.

The table below shows the average price of wheat in European countries between 1838 and 1845, measured relative to the price of silver for comparison across countries, and also the peak price reached during the period of food shortage. There are three groups of countries: those where violent revolutions took place, those in which there was substantial constitutional change in 1848 without widespread violence, and those where no revolution occurred.

1. Explain, using supply and demand curves, how a poor wheat harvest could lead to price rises and food shortages.
2. Find a way to present the data below to show that the size of the price shock (that is, the sudden change in prices), rather than the level of prices, is associated with the likelihood of revolution.
3. Do you think this is a plausible explanation for revolution?
4. In April 2011 a journalist suggested that similar factors may have played a part in the Arab Spring that began in late 2010 in the Middle East and North Africa. Read the blog post by clicking on the link. What do you think of this hypothesis?

		AVE. PRICE 1838-45	MAX. PRICE 1845-48
Violent revolution 1848	AUSTRIA	52.9	104.0
	BADEN	77.0	136.6
	BAVARIA	70.0	127.3
	BOHEMIA	61.5	101.2
	FRANCE	93.8	149.2
	HAMBURG	67.1	108.7
	HESSE-DARMSTADT	76.7	119.7
	HUNGARY	39.0	92.3
	LOMBARDY	88.3	119.1
	MECKLENBURG-SCHWERIN	72.9	110.9
	PAPAL STATES	74.0	105.1
	PRUSSIA	71.2	110.7
	SAXONY	73.3	125.2
	SWITZERLAND	87.9	146.7
	WURTTEMBERG	75.9	128.7

		AVE. PRICE 1838-45	MAX. PRICE 1845-48
Immediate constitutional change 1848	BELGIUM	93.8	140.1
	BREMEN	76.1	109.5
	BRUNSWICK	62.3	100.3
	DENMARK	66.3	81.5
	NETHERLANDS	82.6	136.0
	OLDENBURG	52.1	79.3

		AVE. PRICE 1838-45	MAX. PRICE 1845-48
No revolution 1848	ENGLAND	115.3	134.7
	FINLAND	73.6	73.7
	NORWAY	89.3	119.7
	RUSSIA	50.7	44.1
	SPAIN	105.3	141.3
	SWEDEN	75.8	81.4

Source: *Berger, Helge, and Mark Spoerer. 2001. 'Economic Crises and the European Revolutions of 1848.' The Journal of Economic History 61 (2): 293–326.*

8.7 ENTRY TO THE MARKET

Another reason for a change in supply in a market is the entry of more firms, or the exit of existing firms. So far in our analysis of the bread market we have just assumed that there are 50 bakeries. But if the profits of the bakeries were above normal profits, so that bakery owners are receiving economic rents, then other firms might want to invest in the baking business. Conversely, if profitability fell—perhaps as a result of a fall in demand—economic profits could become negative, causing some bakeries to close down.

Let's start again from the original equilibrium in the bread market, in which 5,000 loaves are produced, and sold at €2 each. There are 50 bakeries, and we will assume they all have the same costs: the isocost and marginal cost curves are shown in Figure 8.13. Remember that isoprofit curves slope down where the marginal cost is less than the price, because making one more loaf would increase profit unless the price went down, and similarly they slope up where the marginal cost is above the price. Since they are price-takers, each bakery is producing at the point on its own marginal cost curve, where price equals €2, making 100 loaves. The isoprofit curve furthest to the left shows points at which economic profits are zero (price equals marginal cost, and the firm is earning what economists call *normal profits*). You can see that, when price is equal to €2 and quantity is equal to 100, the bakery is above this curve at point A—so it is making a positive economic profit.

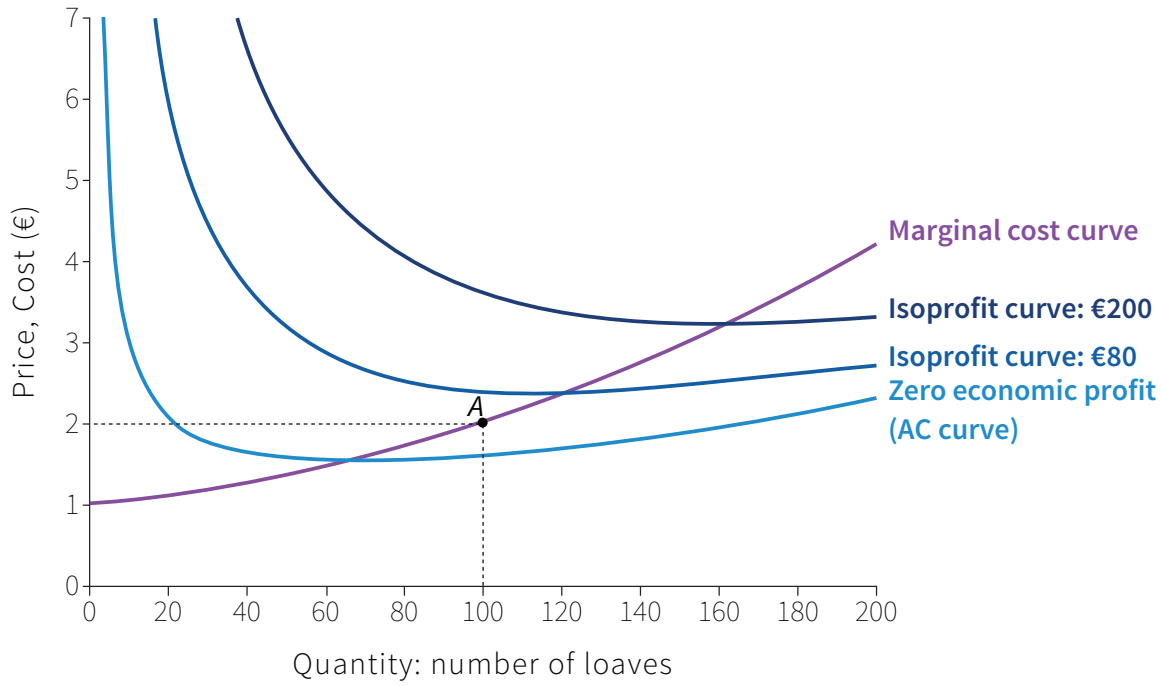


Figure 8.13 *Isoprofit curves and marginal cost curve for the bakery.*

Since there is an opportunity for making greater than normal profit by selling bread in this city, other bakeries may decide to enter the market. There will be some costs of entry—of acquiring and equipping the premises, for example—but provided these are not too high (or if premises and equipment can be easily sold if the venture doesn't work out) it will be worthwhile to do so.

When more bakeries have entered, more bread will be supplied at each level of the market price. Although the reason for the supply increase is different, the effect on the market equilibrium is the same: a fall in price and a rise in bread sales. Figure 8.14 shows the effects on equilibrium of more firms entering the market. The bakeries once again start off at point *A*, selling 5,000 loaves of bread for €2. The entry of new firms shifts out the supply curve. There is more bread for sale at each price, so at the original price there would be excess supply. The new equilibrium is at point *B* with a lower price and higher bread sales.

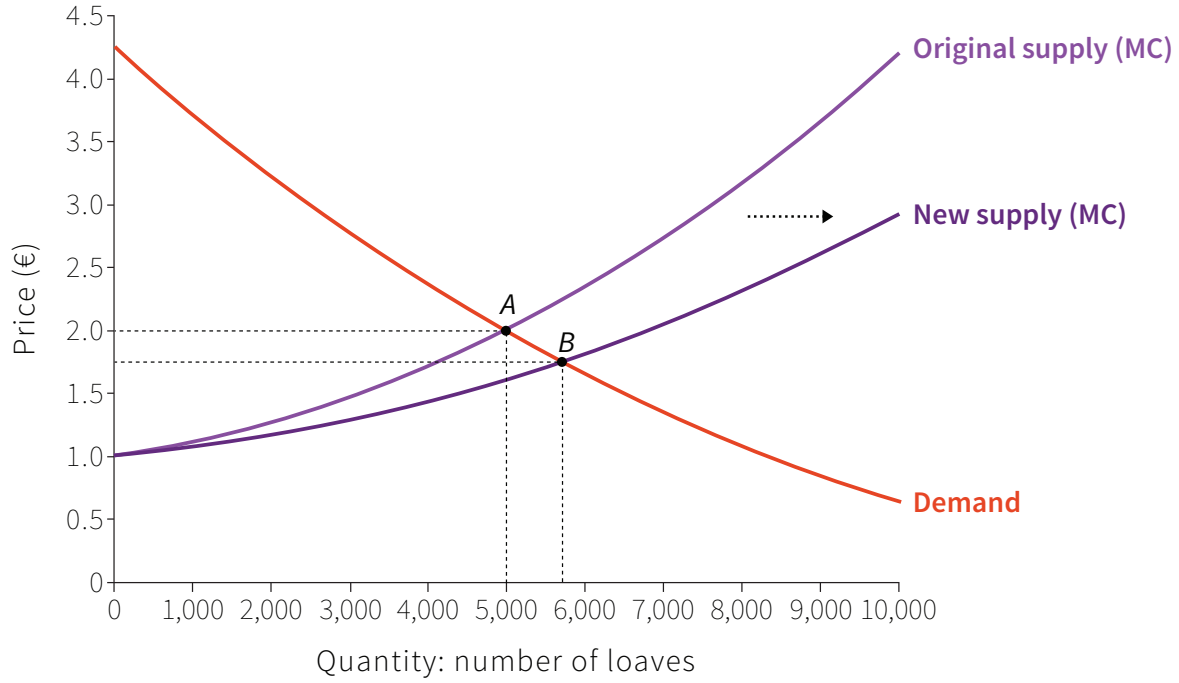


Figure 8.14 *An increase in the supply of bread (more firms enter).*

The entry of new firms is unlikely to be welcomed by the existing bakeries. Their costs have not changed, but the market price has fallen to €1.75, so they must be making less profit. If you look again at Figure 8.13, you will see that they will be on a lower isoprofit curve, producing less output than before. However, they are still above the lightest blue curve, making positive economic profits—perhaps more firms can be expected to enter the market in future.

The original bread market equilibrium at point A in Figure 8.14 is described as a *short-run equilibrium*. The phrase "short-run" is used to indicate that we are holding something constant. In this case, we mean that point A is the equilibrium while the number of firms in the market remains constant. In the longer run, firms may leave

or enter the market, leading to a change in market supply. Closing down or opening new firms takes time, so this cannot happen instantaneously. In general we expect more firms to enter if profits are high. Similarly, if a fall in demand leads to losses, firms eventually leave.

In the long run we would expect the number of firms in the market to be such that no more than normal profits could be made by entering the market. Profits to be made in the bread market would be no higher than the profits potential bakery owners could make by using their assets elsewhere. And, if any bakery owners could do better by putting their premises to a different use (or by selling them and investing in a different business) we would expect them to do so. Although no one would be earning more than normal profits, no one should be earning less than normal profits either.

As long as bakeries are making positive economic profits, firms would continue to enter, increasing supply and lowering the market price, until the price of a loaf of bread was equal to the average cost of producing it (including the opportunity cost of capital). Figure 8.15 shows how the market for bread changes from the initial short-run equilibrium at point A, through entry of new firms, to a long run equilibrium in which the bakeries are making normal profits. The left-hand panel shows the marginal cost curve for each bakery, and the right-hand panel shows the supply and demand curves for the market.

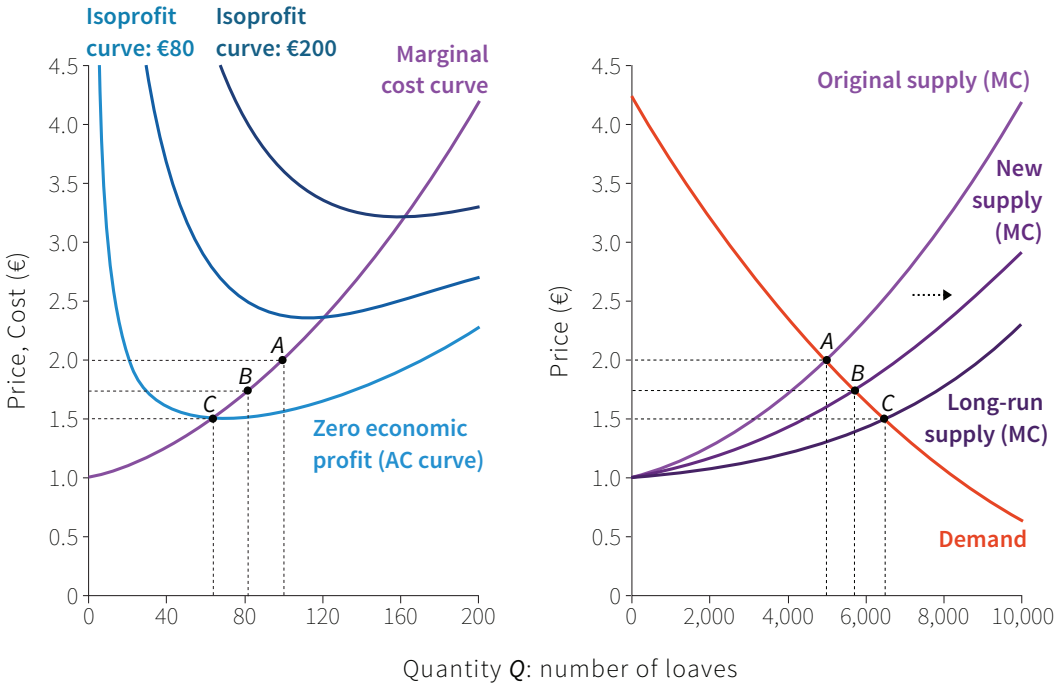


Figure 8.15 *The market for bread in the short run and the long run.*

Initially there are 50 bakeries. The market is at a short-run equilibrium at point *A*. The price of a loaf of bread is €2, and the bakeries' profits are above the normal level. They are earning rents, so more bakeries will wish to enter. When new firms enter the supply curve shifts to the right. The new equilibrium is at point *B*. The price has fallen to €1.75. There are more bakeries selling more bread in total, but each one is producing less than before and making less profit. At *B*, the price is still above the average cost—bakeries are making greater-than-normal profits. This is still only a short-run equilibrium, because more will want to enter. More bakeries will enter, lowering the market price, until the price is equal to the average cost of a loaf, and bakeries are making normal profits. The long-run equilibrium is at point *C*.

In our model, in which we are assuming that all bakeries have the same cost functions, the long run equilibrium will be reached when the price is exactly €1.52 and each bakery is producing 66 loaves. This is point C at which the marginal cost curve cuts the average cost curve. When this point is reached the price of bread is equal to both the marginal and the average cost ($P = MC = AC$), and every bakery's economic profit is zero.

We can work out using Figure 8.15 how many bakeries there will be in the long-run equilibrium. We know from the left-hand panel that the price must be €1.52, because that is the point on the firm's supply curve where the firm makes normal profits ($P = MC = AC$), and each bakery produces 66 loaves. From the demand curve in the right-hand panel we can deduce that at this price the quantity of bread sold will be 6,500 loaves. So the number of bakeries in the market must be $6{,}500/66 = 98$.

Note that for price-taking firms it is always true that *price* = *MC*, both in the short run and the long run. In the long run, it must also be true that *price* = *AC*; otherwise another firm would enter the market. In our model, *price* = *AC* for all of the bakeries, because they all have the same cost function. That would not be true if some firms had better bread-making technologies than others. In that case, the AC of the firms with lowest costs would be below the price, and they would earn rents from their superior technology. The marginal firm in the long-run equilibrium—the one just indifferent between entering the market and staying out—would have $P = AC$. The AC of the higher cost firms would be above the market price, so they would stay out. So:

- Firms will continue to enter a market *until it is not possible to make more than normal profits.*
- In the long-run competitive equilibrium with market price P^*, *the marginal cost of every firm will be equal to the price* ($P^* = MC$).
- *The average cost of the marginal firm will be equal to the price* ($P^* = AC$).

DISCUSS 8.6: THE MARKET FOR QUINOA

Consider again the market for quinoa. The changes shown in Figure 8.10 can be analysed as shifts in demand and supply.

1. Suppose there was an unexpected increase in demand for quinoa in the early 2000s (a shift in the demand curve). What would you expect to happen to the price and quantity initially?
2. Assuming that demand continued to rise over the next few years, how do you think farmers responded?
3. Why did the price stay constant until 2007?
4. How could you account for the rapid price rise in 2008 and 2009?
5. Would you expect the price to fall eventually to its original level?

The graphs in Figure 8.10 are taken from a World Bank blog that tells you more about quinoa.

8.8 THE EFFECTS OF TAXES

Taxation can be used by governments to raise revenue (to finance government spending, or redistribute resources) or to affect the allocation of goods and services in other ways—perhaps because the government considers a particular good is harmful. The supply and demand model is a useful tool for analysing the effects of taxation.

Using taxes to raise revenue

Revenue-raising through taxation has a long history. Take the taxation of salt, for example. For most of history salt was used all over the world as a preservative, allowing food to be stored, transported and traded. The ancient Chinese advocated taxing it—since people needed it, however high the price. Salt taxes were used in ancient India, and in Europe by medieval kings; they were an effective but often resented tool for ruling elites. Resentment of high salt taxes played an important part in the French Revolution, and Gandhi led protests against the salt tax imposed by the British in India.

Figure 8.16 illustrates how a tax on salt might work. Initially the market equilibrium is at point *A*; the price is *P**, and the quantity of salt traded is *Q**. Suppose that a sales tax of 30% is imposed on the price of salt, to be paid to the government by the suppliers. If suppliers have to pay a 30% tax, their marginal cost for each unit of salt increases by 30%. So the supply curve shifts: it is 30% higher at each quantity.

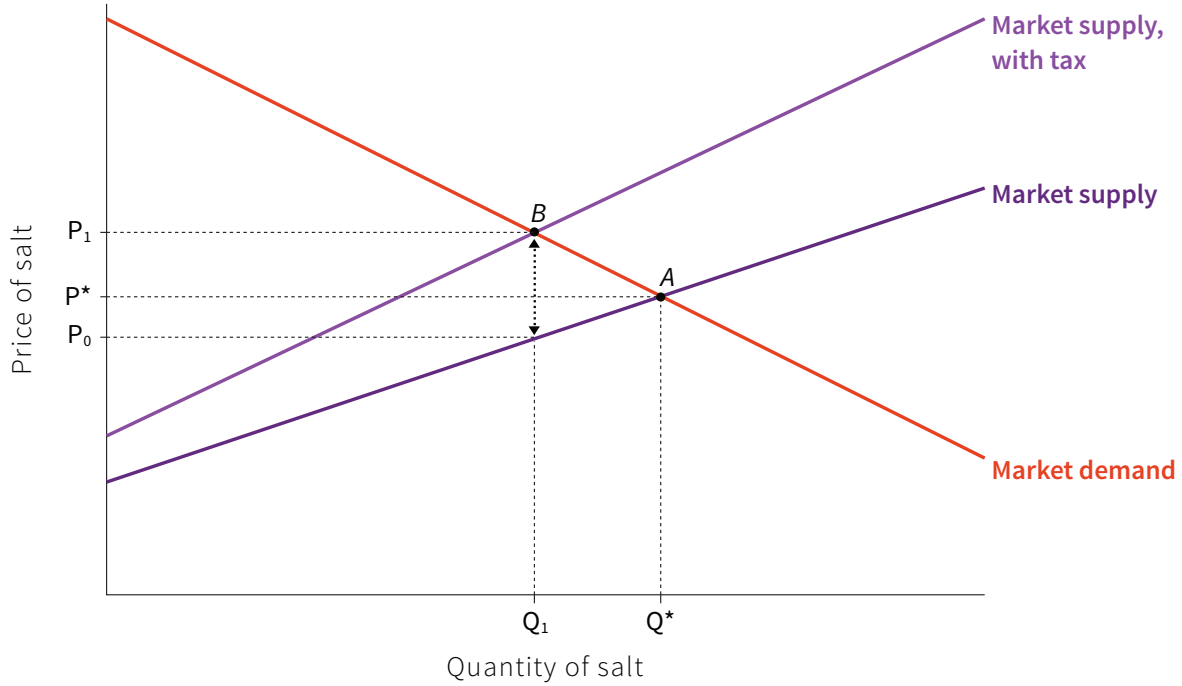


Figure 8.16 *The effect of a 30% salt tax.*

The price received by suppliers (after they have paid the tax) is P_0. The double-headed arrow shows the tax paid to the government on each unit of salt sold.

The new equilibrium is at point *B*, where a lower quantity of salt is traded. Although the consumer price has risen, note that it is not 30% higher than before. The price paid by consumers, P_1, is 30% higher than the price received by the suppliers, net of the tax, which is P_0. Suppliers receive a lower price than before; they produce less, and their profits will be lower. This illustrates an important feature of taxes: it is not necessarily the payer of the tax who feels its main effect. In this case, although the tax is paid by suppliers, the incidence of the tax falls partly on consumers and partly on producers.

Figure 8.17 shows the effect of the tax on consumer and producer surplus:

- *Consumer surplus falls*: Consumers pay a higher price, and less salt is sold.
- *Producer surplus falls*: They produce less, and receive a lower net price.
- *Total surplus is lower*: Even taking account of the tax revenue received by the government, the tax causes a deadweight loss.

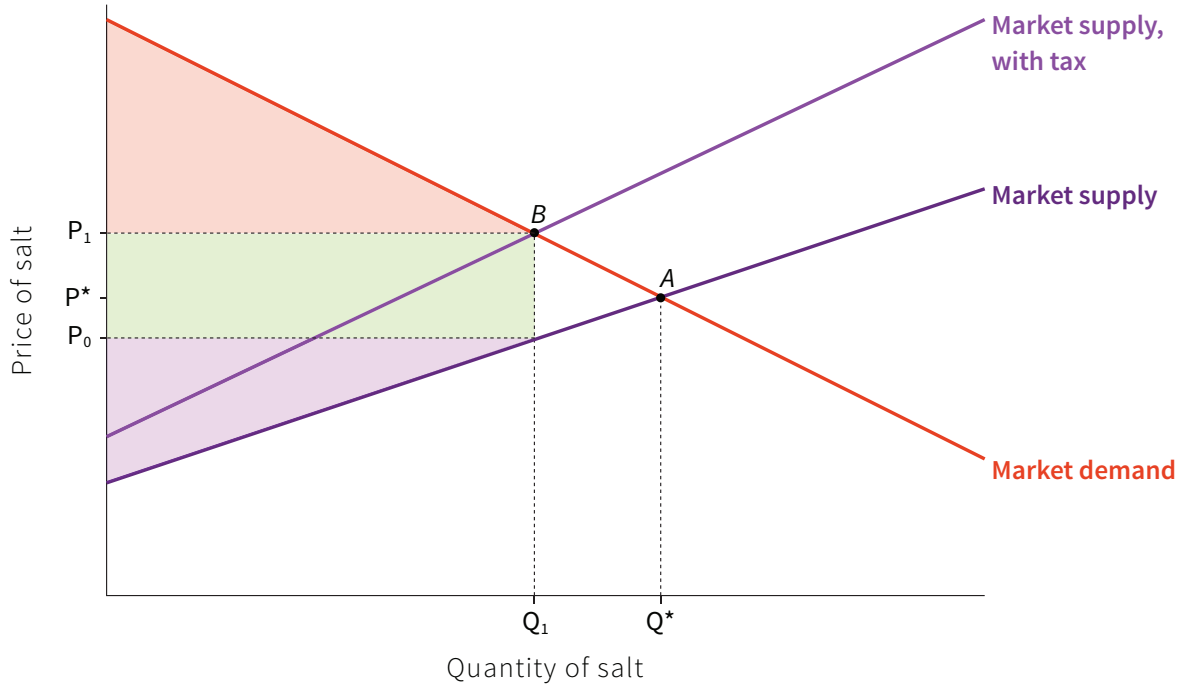


Figure 8.17 *Taxation and deadweight loss.*

A tax equal to $(P_1 - P_0)$ is paid on each of the Q_1 units of salt that are sold. The green area is the total tax revenue. There is a deadweight loss equal to the area of the white triangle.

When the salt tax is imposed, the total surplus from trade in the salt market is given by:

$$total\ surplus = consumer\ surplus + producer\ surplus + government\ revenue$$

Since the quantity of salt traded is no longer at the level that maximises gains from trade, the tax has led to a deadweight loss.

In general, taxes change prices, and prices change buyers' and sellers' decisions, which can cause deadweight loss. To raise as much revenue as possible the government would prefer to tax a good for which demand is not very responsive to price, so that the fall in quantity traded is quite small—that is to say, a good with a low elasticity of demand. That is why the ancient Chinese recommended salt as a suitable commodity to tax.

We can think of these three components together as a measure of the welfare of society as a whole (although this depends on whether the tax revenue is to be used for the benefit of society), so there is a second reason for a government that cares

about welfare to prefer taxing goods with low demand elasticity—the loss of total surplus will be lower. Whether this is a good thing, of course, depends on what the government does with the revenues that it collects:

- If the government spends the revenue providing basic goods and services that enhance the wellbeing of the population, the tax and resulting expenditure may enhance public welfare—even though it reduces consumer surplus in the particular market that is taxed.
- If the government spends the revenues on some activity that does not contribute to the citizens' well being, then the lost consumer surplus is just a reduction in their living standard.

Therefore taxes can improve or reduce overall welfare. The most that we can say is that taxing a good whose demand is inelastic is an efficient way to transfer the surplus from consumers to the government.

The effect of a tax on the market equilibrium is in some ways similar to the effect of a price-setting firm selling a differentiated good: the firm uses its market power to raise the price in pursuit of economic rent; the government uses its power to levy taxes to raise the price and collect revenue. Both have the effect of reducing the quantity sold. The government's power to levy taxes depends on the institutions it can use to enforce and collect them.

One of the reasons for the use of salt taxes in earlier times was that it was relatively easy for a powerful ruler to take full control of salt production, in some cases as a monopolist. In the notorious case of the French salt tax, the monarchy not only controlled all salt production; it also forced its subjects to buy up to 7kg of salt each per year.

In March and April 1930 the artificially high price of salt in British colonial India sparked one of the defining moments of the Indian independence movement: Mahatma Gandhi's salt march to acquire salt from the Indian ocean. Similarly in what came to be called the *Boston tea party* in 1773 American colonists objecting to a British colonial tax on tea dumped a cargo of tea into the Boston harbour.

Resistance to taxes on inelastic goods arises for the very reason they are imposed: they are difficult to escape!

In many modern economies the institutions for tax collection are well-established, usually with democratic consent. Provided that citizens perceive taxes have been implemented fairly, using them to raise revenue is accepted as a necessary part of social and economic policy. We will now look at another reason why governments may decide to levy taxes.

Using taxes to change behaviour

Policymakers in many countries are interested in the idea of using taxes to deter consumption of unhealthy foods with the objective of improving public health and tackling the obesity epidemic. In Unit 7 we looked at some data and estimates of demand elasticities for food products in the United States, which help to predict how higher prices might affect people's diets there. Some countries have already introduced food taxes. Finland has a "sweet tax" on sweets, ice cream and soft drinks. Several, including France, Norway, Mexico, Samoa and Fiji, tax sweetened drinks. Hungary's "chips tax" is aimed at products carrying proven health risks, particularly those with high sugar or salt content. In 2011, the Danish government introduced a tax on products with high saturated fat content.

The level of the Danish tax was 16 Danish kroner (kr) per kilogram of saturated fat, corresponding to 10.4kr per kg of butter. Note that this was a *specific tax*, levied as a fixed amount per unit of butter. A tax like the one we analysed for salt, levied as a percentage of the price, is known as an *ad valorem tax*. According to a study of the Danish fat tax, it corresponded to about 22% of the average butter price in the year before the tax. The study found that it reduced the consumption of butter and related products (butter blends, margarine, and oil) by between 15% and 20%. We can illustrate the effects in the same way as we did for the salt tax, using the supply and demand model. (We are assuming here that butter retailers are price-takers.)

Figure 8.18 shows a demand curve for butter, measured in kilograms per person per year. The numbers correspond approximately to Denmark's experience. We have drawn the supply curve for butter as almost flat, on the assumption that the marginal cost of butter for retailers does not change very much as quantity varies. The initial equilibrium is at point A, where the price of butter is 45kr per kg, and each person consumes 2kg of butter per year.

A tax of 10kr per kg shifts the supply curve upwards and leads to a rise in price to 54kr, and a fall in consumption to 1.6kg. The consumer price rises by 9kr—almost the full amount of the tax—and the suppliers' revenue per kg of butter, net of the tax, falls to 44kr. In this case, although the tax is paid by suppliers, the incidence of the tax is felt mainly by consumers. Of the 10kr tax per kg, the consumer effectively pays 9kr, while the supplier or producer pays 1kr. So the price received by the retailers, net of tax, is only 1kr lower.

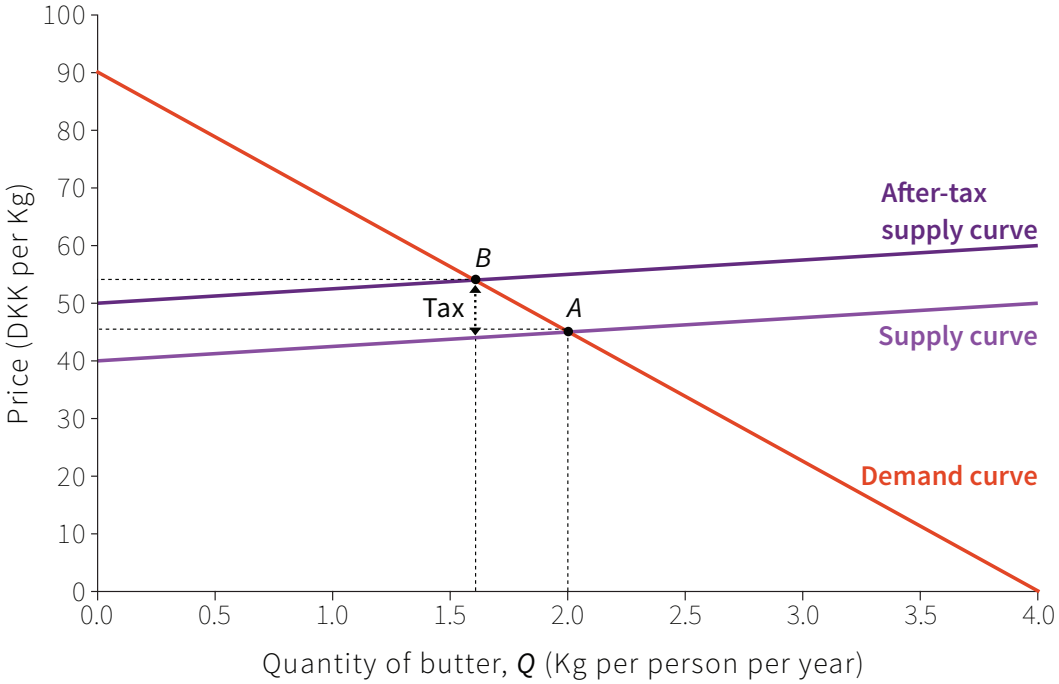


Figure 8.18 *The effect of a fat tax on the retail butter market.*

Figure 8.19 shows what happens to consumer and producer surplus as a result of the fat tax.

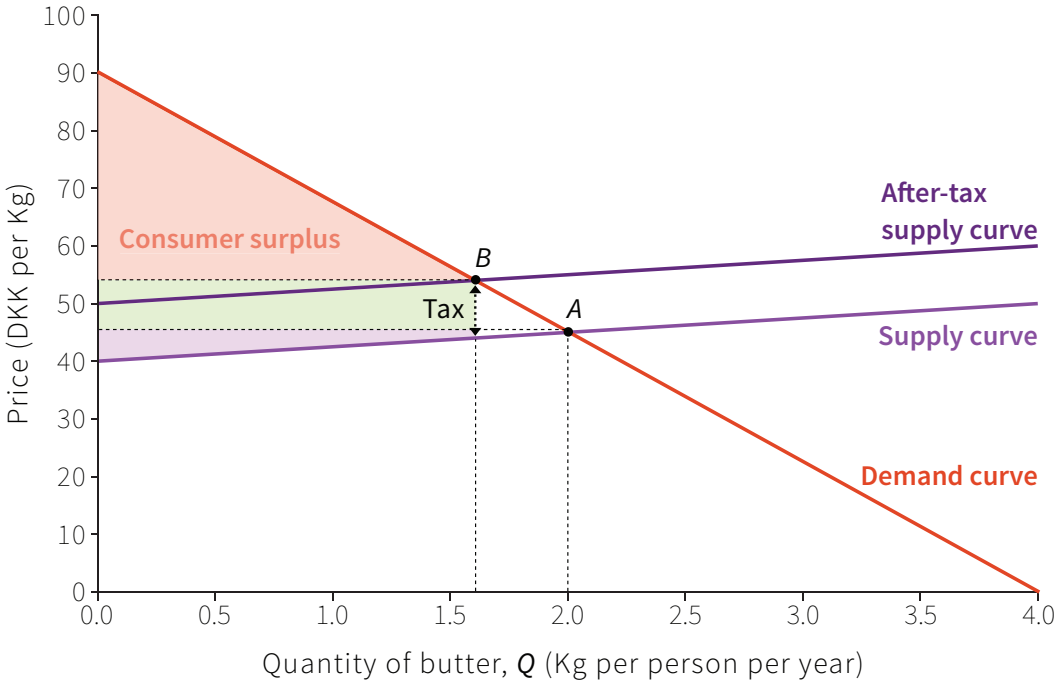


Figure 8.19 *The effect of a fat tax on consumer and producer surplus for butter.*

Again, the consumer and producer surpluses fall. The area of the green rectangle represents the tax revenue; with a tax of 10kr per kg, and equilibrium sales of 1.6 kg per person, tax revenue is *10 x 1.6 = 16kr* per person per year.

How effective was the fat tax policy? We have simplified the analysis here by focusing only on the market for butter. As we discussed in Unit 7, to evaluate the effect of the tax on health we should take into account the cross-price effects—the changes in consumption of other foods caused by the tax. We have also assumed that suppliers are price-takers. The study of the Danish tax allowed for the possibility that some retailers have market power, and for the joint price effects on all butter-related products, taxed at different levels depending on saturated fat content, but not for changes to the rest of the diet. Nevertheless Figures 8.18 and 8.19 illustrate some important implications of the tax:

- *Consumption of butter products fell by 20%*: You can see this in Figure 8.18. In this respect the policy was successful.
- *There was a large fall in surplus—especially consumer surplus*: You can see this in Figure 8.19. The tax had a big effect on quantity, so it is less effective at revenue-raising than a tax on a good with less elastic demand. But recall that the government's aim when it implemented the fat tax policy was not to raise revenue; the intention was to reduce quantity. So the fall in consumer surplus was inevitable. The loss of surplus caused by a tax is a deadweight loss, which sounds negative. But in this case the policymaker might think that this loss is actually a gain if that policymaker considers the "good", butter, to be a "bad" for consumers.

DISCUSS 8.7: THE DEADWEIGHT LOSS OF THE BUTTER TAX

Food taxes such as the ones discussed in Unit 7 and above in this unit are often intended to shift consumption towards a healthier diet, but give rise to deadweight loss.

Why do you think a policymaker and a consumer might interpret this deadweight loss differently?

One aspect of taxation that is not illustrated in our supply-and-demand analysis is the cost associated with collecting it. Although the Danish fat tax was successful in reducing fat consumption, the government abolished it after only 15 months because of the administrative burden it placed on firms. Any taxation system requires effective mechanisms for tax collection, and designing taxes that are simple to administer (and difficult to avoid) is an important consideration for tax policy in general. Policymakers who want to introduce food taxes will need to find ways of

minimising administrative costs. But since the costs cannot be eliminated, they will also need to consider whether the health gain (and reduction of costs of bad health) will be sufficient to offset them.

8.9 THE MODEL OF PERFECT COMPETITION

To apply the model of supply and demand, we have assumed throughout this unit that buyers and sellers are price-takers. In what kinds of markets would we expect to see price-taking on both sides? To generate competition between sellers, and force sellers to act as price-takers, we need:

- *Many undifferentiated sellers*: As Marshall discussed when he introduced the model of supply and demand, there must be many sellers, all selling identical goods. If their goods were differentiated, then each one would have some market power.
- *Many buyers all wanting to buy the good*: Each of whom will choose whichever seller has the lowest price.
- *Buyers know the sellers' prices*: If they do not, they cannot choose the lowest one.

Similarly, for the buyers to be price-takers:

- *There must be many other buyers*: So that sellers have no reason to sell to someone who would pay less than everyone else.
- *Buyers must be competing with other buyers, and sellers with other sellers*: If sellers act as a cartel, for example, they are not price-takers—they can jointly choose the price.

In this market all buyers and sellers are acting competitively, so that they force each other to be price-takers.

PERFECT COMPETITION

A hypothetical market in which:

- The good or service being exchanged is homogeneous; it does not differ from one seller to another
- There are a very large number of potential buyers and sellers of this good, each acting independently of the others
- Buyers and sellers can readily know the prices at which other buyers and sellers are exchanging the good

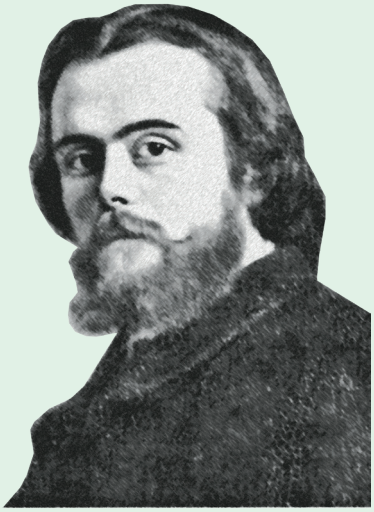
A market with all of these properties is described as *perfectly competitive*. We can predict that the equilibrium in such a market will be a competitive equilibrium—so it will have the following characteristics:

- All transactions take place at a single price. This is known as the *Law of One Price*.
- At that price, the amount supplied equals the amount demanded: the *market clears*.
- No buyer or seller can benefit by altering the price they are demanding or offering. They are *price-takers*.
- All potential *gains from trade* are realised.

Léon Walras, a 19th-century French economist, built a mathematical model of an economy in which all buyers and sellers are price-takers that has been influential in how many economists think about markets.

GREAT ECONOMISTS

LÉON WALRAS



Léon Walras (1834-1910) was a founder of the neoclassical school of economics. He was an indifferent student, and twice failed the entrance exam to the École Polytechnique in Paris, one of the most prestigious universities in his native France. He studied engineering at the School of Mines instead. Eventually his father, an economist, convinced him to take up the challenge of making economics into a science.

The pure economic science to which he aspired was the study of relationships among things, not people, and he had notable success in eliminating human relationships from his modelling. "The pure theory of economics," he wrote, "resembles the physico-mathematical sciences in every respect."

His device for simplifying the economy so that it could be expressed mathematically was to represent interactions among economic agents as if they were relationships among inputs and outputs, and to focus entirely on the economy in equilibrium. In the process the entrepreneur, a key actor in wealth creation from the Industrial Revolution to today, simply disappeared from Walrasian economics:

"Assuming equilibrium, we may even go so far as to abstract from entrepreneurs and simply consider the productive services as being, in a certain sense, exchanged directly for one another... "
— Léon Walras, *Elements of Theoretical Economics* (1874)

Walras represented basic economic relationships as equations, which he used to study the workings of an entire economy composed of many interlinked markets. Prior to Walras most economists had considered these markets in isolation: they would have studied, for example, how the price of textiles is determined on the cloth market, or land rents on the land market.

A century before Walras, a group of French economists called the physiocrats had studied the circulation of goods throughout the economy, as if the flow of goods from one sector to another in the economy was comparable to the circulation of blood in the human body (one of the leading physiocrats was a medical doctor). But the physiocrats' model was little more than a metaphor that drew attention to the interconnectedness of markets.

Walras used mathematics, rather than medical analogies, to create was what is now called *general equilibrium theory*, a mathematical model of an entire economy in which all buyers and sellers act as price-takers and supply equals demand in all markets. Walras' work was the basis of the proof, much later, of the *invisible hand theorem*, giving the conditions under which such an equilibrium is Pareto efficient. (The invisible hand game in Unit 4 is an example of the conditions in which the pursuit of self-interest can benefit everyone.)

Walras had defended the right to private property, but to help the working poor he also advocated the nationalisation of land and the elimination of taxes on wages.

Seven years after his death the general equilibrium model was to play an important role in the debate about the feasibility and desirability of centralised economic planning compared to a market economy. In 1917 the Bolshevik Revolution in Russia put the economics of socialism and central planning on the agenda of many economists but, surprisingly, it was the defenders of central planning, not the advocates of the market, who used Walras' insights to make their points.

Friedrich Hayek, and other defenders of capitalism, criticised the Walrasian general equilibrium model. Their argument: by deliberately ignoring the fact that a capitalist economy is constantly changing, and therefore not taking into account the contribution of entrepreneurship and creativity in market competition, Walras had missed the true virtues of the market.

The model of perfect competition describes an idealised market structure in which we can be confident that the assumption of price-taking that underlies our model of supply and demand will hold. Markets for agricultural products such as wheat, rice, coffee, or tomatoes look rather like this, although goods are not truly identical, and it is unlikely that everyone is aware of all the prices at which trade takes place. But it is nevertheless clear that they have very little, if any, power to affect the price at which they trade: they are price-takers.

In other cases—for example, markets where there are some differences in the quality of goods—there may still be enough competition that we can assume price-taking, in order to obtain a simple model of how the market works. A simplified model can provide useful predictions when the assumptions underlying it are only approximately true. Judging when it is appropriate to draw conclusions about the real world from a simplified model is an important skill of economic analysis.

For example, we know that markets are not perfectly competitive when products are differentiated. Consumers' preferences differ, and we saw in Unit 7 that firms have an incentive to differentiate their product, if they can, rather than to supply a product similar or identical to others. Nevertheless, the model of supply and demand can be a useful approximation to help us to understand how some markets for non-identical products behave.

Figure 8.20a shows the market for an imaginary product which we have called *Choccos*, for which there are close substitutes: many similar products compete in the wider market for chocolate bars. Due to competition from other chocolate bars, the demand curve is almost flat. The range of feasible prices for Choccos is narrow, and the firm chooses a price and quantity where the marginal cost is close to the price. So this firm is in a similar situation to a firm in a perfectly competitive market (although, surprisingly, the market for chocolate bars in the real world may be yet another example of imperfect competition). We can construct a supply curve for the chocolate bar market as a whole from the individual firms' marginal cost curves. The equilibrium price in the market for chocolate bars then determines the feasible prices for Choccos—they have to be sold at a similar price to other chocolate bars.

The narrow range of feasible prices for this firm is determined by the behaviour of its competitors. So the main influence on the price of Choccos is not the firm, but the market for chocolate bars as a whole. Since all the firms will be producing at similar prices, which will be close to their marginal costs, we lose little by ignoring the differences between them and assuming that each firm's supply curve is its marginal cost curve, then finding the equilibrium in the wider market for chocolate bars.

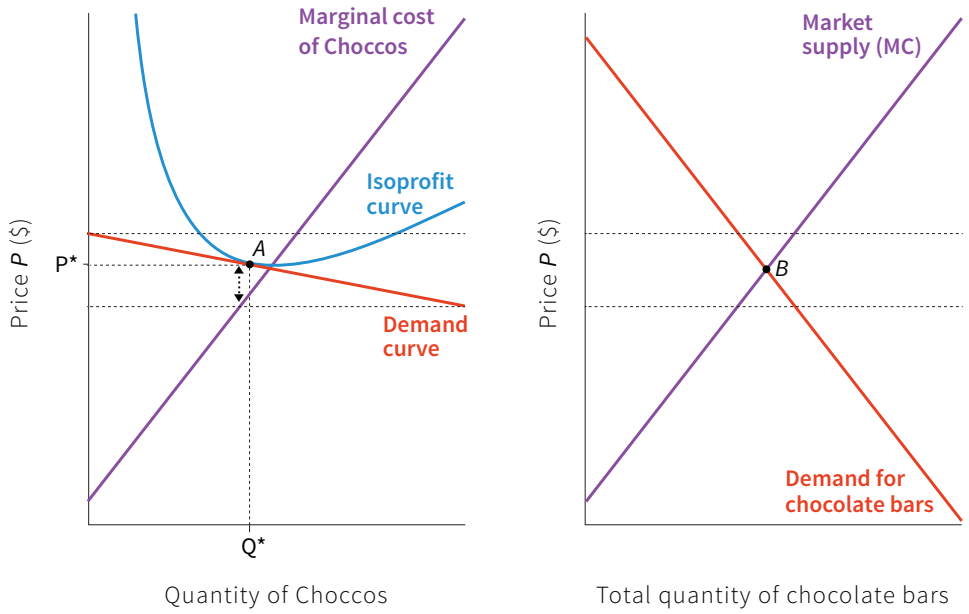


Figure 8.20 *The market for Choccos and chocolate bars.*

The left hand panel shows the market for Choccos, produced by one firm. There are many close substitutes in the wider market for chocolate bars.Due to competition, the demand curve for Choccos is almost flat. The range of feasible prices is narrow. The firm chooses a price P^* similar to its competitors, and a quantity where MC is close to P^*. Whatever the price of its competitors, it would produce close to its marginal cost curve. So the firm's MC curve is approximately its supply curve. We can construct the market supply curve for chocolate bars in the right hand panel from the marginal cost curves of all the firms making chocolate bars. If most consumers do not have strong preferences for one firm's product, we can draw a market demand curve for chocolate bars. The equilibrium price in the chocolate bar market (right hand panel) determines the narrow range of prices from which the Chocco firm can choose (left hand panel)—it will have to set a price quite close to that of other chocolate bars.

We have already taken this approach when we analysed the market for butter in Denmark. In practice, it is likely that some retailers who sell butter have some power to set prices. A local shop may be able to set a price that is higher than the price of butter elsewhere, knowing that some shoppers will find it convenient to buy rather than searching for a lower price. However, it is reasonable to assume that they don't have much wiggle room to set prices, and are strongly influenced by the prevailing market price. So price-taking is a good approximation for this market—good enough, at least, that the supply and demand model can help us to understand the impact of a fat tax.

8.10 LOOKING FOR COMPETITIVE EQUILIBRIA

If we look at a market in which conditions seem to favour perfect competition—many buyers and sellers of identical goods, acting independently—how can we tell whether it satisfies the conditions for a competitive equilibrium? Economists have used two tests:

1. Do all trades take place at the same price?
2. Are firms selling goods at a price equal to marginal cost?

The difficulty with the second test is that it is often difficult to measure marginal cost. But Lawrence Ausubel, an economist, was able to do this for the US bank credit card market in the 1980s. At this time 4,000 banks were selling an identical product: credit card loans. The cards were mostly Visa or Mastercard, but the individual banks decided the price of their loans—that is, the interest rate. The banks' cost of funds—the opportunity cost of the money loaned to credit card holders—could be deduced from other interest rates in financial markets. Although there were other components of marginal cost, the cost of funds was the only one that varied substantially over time. If the credit card market were competitive, we would expect to see the interest rate on credit card loans rise and fall with the cost of funds.

Ausubel found that this didn't happen.

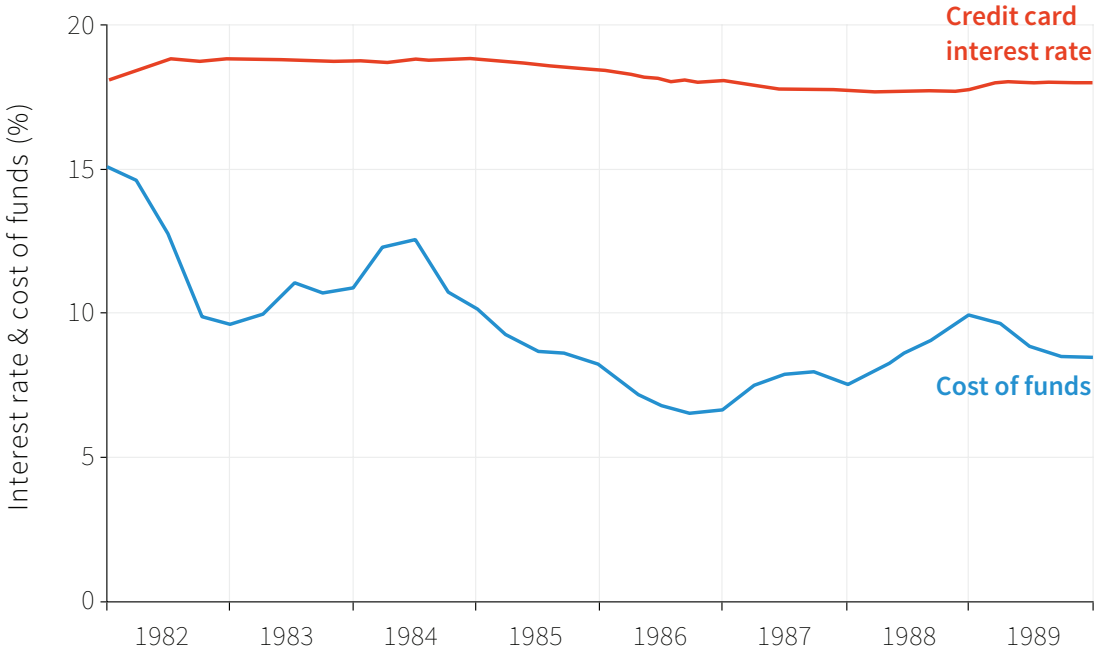


Figure 8.21 *Ausubel's credit card data.*

Source: *Figure 1 in Ausubel, Lawrence M. 1991. 'The Failure of Competition in the Credit Card Market.' American Economic Review 81 (1): 50–81.*

As Figure 8.21 shows, when the cost of funds fell from 15% to below 7%, there seemed to be almost no effect on the price of credit card loans.

Why do the banks not cut their interest rates when their costs fall? He suggested two different possibilities:

- *It may be difficult for consumers to change credit card provider*: In that case the banks are not forced to compete with each other; so they keep prices high when costs fall.
- *Banks might not be able to decide which of their customers are bad risks*: This would be a problem in this market, because the bad risks are most sensitive to prices. The banks do not want to lower their prices for fear of attracting the wrong kind of customer.

Perfect competition requires that consumers are sufficiently sensitive to prices to force firms to compete, and this may not be the case in any market where consumers have to search for products. If it takes time and effort to check prices and inspect products, they may decide to buy as soon as they find something suitable, rather than continue the search for the cheapest. When the internet made online shopping feasible, many economists hypothesised that this would make retail markets more competitive: consumers would easily be able to check the prices of many suppliers before deciding to buy.

But often consumers are not very sensitive to prices, even in this environment. You can test the law of one price in online retail competition for yourself, by checking the prices of a particular product that should be the same wherever you buy it—a book, DVD, or household appliance, for example—and comparing them. Figure 8.22 shows the prices of UK online retailers for a particular DVD in March 2014. The range of prices is high: the most expensive seller is charging 66% more than the cheapest.

DISCUSS 8.8: PRICE DISPERSION

What explanations can you suggest for the differing prices for The Hobbit?

The Hobbit: An Unexpected Journey	
SUPPLIER	**PRICE INCLUDING POSTAGE ($)**
Game	14.99
Amazon UK	15.00
Tesco	15.00
Asda	15.00
Base.com	16.99
Play.com	17.79
Zavvi	17.95
The HUT	18.25
I want one of those	18.25
Hive.com	21.11
MovieMail.com	21.49
Blackwell	24.99

Figure 8.22. *Differing prices for the same DVD, from UK online retailers (March 2014).*

Source: *Websites of UK online retailers shown in figure.*

Kathryn Graddy, an economist who specialises in how prices are set, studied the Fulton Fish Market in Manhattan, an institution that appeared to encourage competition. There were about 35 dealers, with stalls close to each other; customers could easily observe the quantity and quality of fish available and ask several dealers for a price. She recorded details of 2,868 sales of whiting by one dealer, including price, quantity and quality of fish, and characteristics of the buyers.

Of course, prices were not the same for every transaction: quality varied, and fish supplies changed from day to day. But her surprising result was that on average Asian buyers paid about 7% less per pound than white buyers. (All of the dealers were white.) There seemed to be no differences between the transactions with white and Asian buyers that could explain the different prices.

How could this happen? If one dealer was setting high prices for white buyers, why did other dealers not try to attract them to their own stalls by offering a better deal? Graddy's results suggested that that the dealers did have some discretion in price-setting; the price difference arose because the Asian buyers were more price sensitive, and persisted because the two groups of buyers were not aware of the difference.

Watch our interview with Kathryn Graddy to find out how she collected her data, and what she discovered about the model of perfect competition as a result.

The evidence in this section suggests that it is hard to find evidence of perfect competition. Nevertheless, we have seen that the model of perfect competition can be a useful approximation, to help us to understand how some markets for non-identical products behave. Even if the conditions for perfect competition are not all satisfied, the model of supply and demand that we have developed in this unit is a very useful tool for economic analysis, applicable when there is sufficient competition that it is reasonable to assume price-taking behaviour.

DISCUSS 8.9: RESTAURANTS IN CHINA

This market research report describes the "full-service" restaurant industry in China.

With the help of this information, discuss whether you would you expect restaurants to produce at a point where their marginal cost is close to the price they receive for a meal.

8.11 CONCLUSION

Looking back over Units 7 and 8 we now have two different models of how firms behave. In the Unit 7 model the firm produces a product that is different from the products of other firms, giving it market power—the power to set its own price. This model applies to the extreme case of a monopolist, who has no competitors at all; common examples are water supply companies, and national airlines with exclusive rights granted by the government to operate domestic flights.

The price-setting model also applies to firm producing differentiated products such as breakfast cereals, cars, or chocolate bars—similar, but not identical, to those of their competitors. In this case the firm still has the power to set its own price, although if it has close competitors demand will be quite elastic and the range of feasible prices narrow.

In the supply and demand model developed in this unit, firms are price-takers. Competition from other firms producing identical products means that they have no power to set their own prices. This model can be useful as an approximate

description of a market in which there are many firms selling very similar products, even if the idealised conditions for a perfectly competitive market do not hold.

In practice economies are a mixture of more competitive markets, and less competitive ones in which firms have more power to set prices—market power. But in some respects firms act the same whether they are the single seller of a good or one of a great many competitors. Most important among their similarities, all firms decide how much to produce, which technologies to use, how many people to hire, and how much to pay them so as to maximise their profits.

CONCEPTS INTRODUCED IN UNIT 8

Before you move on, review these definitions:

- *Price-taking firms*
- *Competitive equilibrium*
- *Exogenous shocks*
- *Entry and exit of firms*
- *Taxation*
- *Model of perfect competition*

But there are important differences. Look back at the decisions made by price-setting firms to maximise profits (Figure 7.2 in Unit 7). Firms in more competitive markets lack either the incentive or the opportunity to do many of these things.

A firm with a unique product will advertise (*Buy Nike!*) to shift the demand curve for its product to the right. But why would a single competitive firm advertise (*Drink milk!*)? This would shift the demand curve for all of the firms in the industry. Advertising in a competitive market is a public good: the firm pays the cost and the benefits go to all of the firms in the industry. If you see a message like "Drink milk!" it is probably paid for by an association of dairies, not by a particular one.

The same is true of expenditures to influence public policy. If a large firm with market power is successful, for example, in relaxing environmental regulations, then it will benefit directly. But lobbying, contributing money to electoral campaigns and other expenses of this type will be unattractive to the competitive firm because the result (a more profit-friendly policy) is a public good.

Similarly, investment in developing new technologies is likely to be undertaken by firms facing little competition, because if they are successful in finding a profitable innovation, the benefits will not be shared with other firms in the industry. The extra profits they will make by a successful innovation are less likely to be competed away by those who copy the innovator. However, successful large firms can emerge by breaking away from the competition and innovating with a new product. The UK's largest organic dairy, Yeo Valley, was once an ordinary farm selling milk, just like thousands of others. In 1994 it established an organic brand, creating new products for which it could charge premium prices. With the help of imaginative marketing campaigns it has grown into a company with 1,400 employees and 65% of the UK organic market.

Figure 8.23 summarises the differences between price-setting and price-taking firms:

PRICE-SETTING FIRM OR MONOPOLY	FIRM IN A PERFECTLY COMPETITIVE MARKET
Sets price and quantity to maximise profits ("price-maker")	Takes market determined price as given and chooses quantity to maximise profits ("price-taker")
Chooses an output level at which marginal cost is less than price	Chooses an output level at which marginal cost equals price
Deadweight losses (Pareto inefficient)	No deadweight losses for consumers and firms (can be Pareto efficient if no-one else in the economy is affected)
Owners receive economic rents (profits greater than normal profits)	If the owners receive any economic rents, the rents are likely to disappear as more firms enter the market
Firms advertise their unique product	Little advertising: it costs the firm, but benefits all firms (it's a public good)
Firms may spend money to influence elections, legislation and regulation	Little expenditure by individual firms on this (same as advertising)
Firms invest in research and innovation; seek to prevent copying	Little incentive for innovation; others would copy (unless the firm can succeed in differentiating its product and escaping from the competitive market)

Figure 8.23. *Price-setting and competitive firms.*

Key points in Unit 8

Price-takers

In a market with many buyers and sellers, individuals and firms may have little influence on prices, due to competition. Then they are called price-takers.

The market-clearing price

A market is in competitive equilibrium if all buyers and sellers are price-takers, and at the prevailing market price, the quantity supplied is equal to the quantity demanded (the market clears).

Competitive equilibrium maximises gains from trade

A competitive equilibrium allocation exploits all possible gains from trade.

The model of perfect competition

The model of a perfectly competitive market describes a set of idealised conditions in which we would expect a competitive equilibrium to occur.

Markets are rarely perfectly competitive

Most markets for real goods don't conform exactly to the model of perfect competition. But price-taking can be a useful approximation, enabling us to use the model of supply and demand as a tool for understanding market outcome

Price-taking firms cannot set a profit-maximising price

Both price-taking and price-setting firms seek to maximise profits, but the former are restricted in the ways they can pursue this objective.

8.13 EINSTEIN

Total surplus and WTP

However the market works, and whatever prices are paid, we can calculate the consumer surplus by adding together the difference between WTP and price paid of all the people who buy, and the producer surplus by adding together the difference between price received and marginal cost of every unit of output:

$$\textit{consumer surplus} = \textit{sum of WTPs} - \textit{sum of prices paid}$$

$$\textit{producer surplus} = \textit{sum of prices received} - \textit{sum of MCs of each unit}$$

Then when we calculate the total surplus, the prices paid and received cancel out:

$$\textit{total surplus} = \textit{sum of WTPs of consumers} - \textit{sum of MCs of producers}$$

When buyers and sellers are price-takers, and the price equalises supply and demand, the total surplus is as high as possible, because the consumers with the highest WTPs buy, and the units of output with the lowest marginal costs are sold. Every trade involves a buyer with higher WTP than the seller's reservation value, so the surplus would go down if we omitted any of them. And if we tried to include any more units of output in this calculation, the surplus would also go down because the WTPs would be lower than the MCs.

8.14 READ MORE

Bibliography

1. Ausubel, Lawrence M. 1991. 'The Failure of Competition in the Credit Card Market.' *American Economic Review* 81 (1): 50–81.
2. Berger, Helge, and Mark Spoerer. 2001. 'Economic Crises and the European Revolutions of 1848.' *The Journal of Economic History* 61 (2): 293–326.
3. Eisen, Michael. 2011. 'Amazon's $23,698,655.93 Book about Flies.' *It Is NOT Junk*. April 22.
4. Ellison, Glenn, and Sara Fisher Ellison. 2005. 'Lessons About Markets from the Internet.' *Journal of Economic Perspectives* 19 (2): 139.

5. Food and Agriculture Organization of the United Nations. 2015. 'FAOSTAT Database.' Accessed July.
6. Graddy, Kathryn. 1995. 'Testing for Imperfect Competition at the Fulton Fish Market.' *The RAND Journal of Economics* 26 (1): 75–92.
7. Graddy, Kathryn. 2006. 'Markets: The Fulton Fish Market.' *Journal of Economic Perspectives* 20 (2): 207–20.
8. IBISWorld. 2015. 'Full-Service Restaurants in China Market Research.' Accessed July.
9. Jensen, Jørgen Dejgård, and Sinne Smed. 2013. 'The Danish Tax on Saturated Fat – Short Run Effects on Consumption, Substitution Patterns and Consumer Prices of Fats.' *Food Policy* 42: 18–31.
10. Marshall, Alfred. (1890) 1920. *Principles of Economics*. 8th ed. London: Macmillan and Co.
11. Mason, Paul. 2011. 'Revolutions and the Price of Bread: 1848 and Now.' *BBC*. April 21.
12. Reyes, Jose Daniel, and Julia Oliver. 2013. 'Quinoa: The Little Cereal That Could.' *The Trade Post*. World Bank. November 22.
13. Seabright, Paul. 2010. *The Company of Strangers: A Natural History of Economic Life* (Revised Edition). Princeton, NJ: Princeton University Press. Chapter 1 is free to read.
14. Stucke, Maurice. 2013. 'Is Competition Always Good?' *OUPblog*. March 25.
15. *The Economist*. 2001. 'Is Santa a Deadweight Loss?' December 20.
16. Waldfogel, Joel. 1993. 'The Deadweight Loss of Christmas.' *American Economic Review* 83 (5).
17. Walras, Leon. (1874) 2014. *Elements of Theoretical Economics: Or the Theory of Social Wealth*. Cambridge: Cambridge University Press.

9

MARKET DISEQUILIBRIUM, RENT-SEEKING AND PRICE-SETTING

Photo: *Alexander Bustos Concha*

HOW PRICES CHANGE, AND HOW MARKETS FOR LABOUR AND FINANCIAL ASSETS WORK

- People take advantage of rent-seeking opportunities when competitive markets are not in equilibrium, often eventually equating supply to demand
- Excess supply—unemployment—is a feature of labour markets even in equilibrium
- Prices are determined in financial markets by trading mechanisms and can change from minute to minute in response to information and beliefs
- Price bubbles can occur, for example in markets for financial assets
- Governments and firms sometimes set prices and adopt other policies so that markets do not clear
- Economic rents help explain how markets work

See www.core-econ.org for the full interactive version of *The Economy* by The CORE Project. Guide yourself through key concepts with clickable figures, test your understanding with multiple choice questions, look up key terms in the glossary, read full mathematical derivations in the Leibniz supplements, watch economists explain their work in Economists in Action – and much more.

Institute for New Economic Thinking
UCL SciencesPo Azim Premji University
FRIENDS PROVIDENT Foundation

Fish and fishing are a major part of the life of the people of Kerala in India; most of them eat fish at least once a day and more than a million people are involved in fishing. But prior to 1997 prices were high and fishing profits limited due to a combination of waste and the bargaining power of fish merchants, who purchased the fishermen's catch and sold it to consumers.

When returning to port to sell their daily catch of sardines to the fish merchants, many fishermen found that the merchants already had as many fish as they needed that day. The fisherman would be forced to dump their worthless catch back into the sea. A lucky few returned to the right port at the right time when demand exceeded supply, and they were rewarded by extraordinarily high prices.

On 14 January 1997, for example, 11 boatloads of fish brought to the market at the town of Badagara found the market oversupplied, and jettisoned their catch. But at fish markets within 15km of Badagara there was excess demand, and 27 buyers would leave the market unable to purchase fish at any price. The luck, or lack of it, of fishermen returning to the ports along the Kerala coast is illustrated in Figure 9.1.

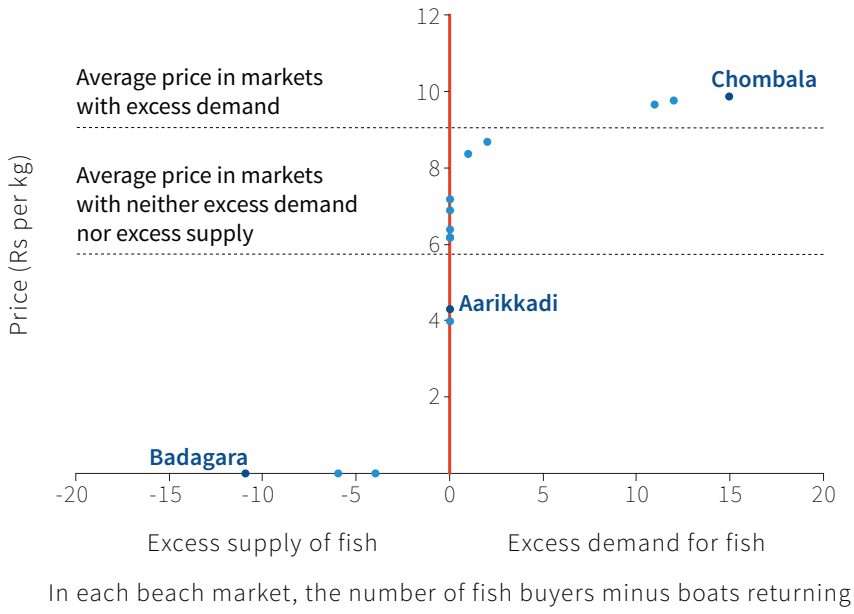


Figure 9.1 *Bargaining power and prices in the Kerala wholesale fish market (14 January 1997). Note two markets had the same outcome of zero excess demand or supply, and a price of Rs4.*

Source: *Jensen, Robert. 2007. 'The Digital Provide: Information (Technology), Market Performance, and Welfare in the South Indian Fisheries Sector.' The Quarterly Journal of Economics 122 (3): 879–924.*

In five markets demand exceeded supply and prices were high. Only seven of the 15 markets did not suffer either from oversupply or buyers unable to purchase what they wanted. In these seven villages prices ranged from Rs4 per kg to more than Rs7 per kg. This is an example of how the Law of One Price—a characteristic of a competitive market equilibrium—is sometimes a poor guide to how actual markets function.

Look more carefully at Figure 9.1. The number of boats arriving at a beach market to find no buyers at any price and throwing their catch into the sea is the measure of excess supply. For example, at the beach market of Badagara, the excess supply is 11 boats arriving (supply) minus zero buyers (demand). Excess demand is measured by the number of fish buyers who could not buy the number of fish they wanted. For example, at Chombala market, 15 buyers reported leaving the market without being able to buy enough fish.

On the morning of 14 January 1997, 21 boats (the three dots to the left of zero) failed to sell their catch; the dead sardines were thrown back into the sea.

When the fishermen have bargaining power because there is excess demand, they get much higher prices. In the seven markets (on the vertical line) with neither excess demand nor excess supply, the average price was Rs5.9 per kg, shown by the horizontal dashed line. For the five markets with excess demand the average was Rs9.3 per kg. The fishermen fortunate enough to have put in at these markets received a price that translates into extraordinary profits, if we assume that the price in the markets with neither excess demand nor supply gave the fishermen an economic profit rate. Of course the following day they may have been the unlucky ones who found no buyers at all, and so would dump their catch into the sea.

This all changed when the fishermen got mobile phones. While still at sea, the returning fishermen would phone the beach fish markets and pick the one at which the prices that day were highest. If they returned to a high-priced market they would earn an economic rent (remember, this is income in excess of their next best alternative, which would be returning to a market with no excess demand or even one with excess supply).

By gaining access to real-time market information (using a phone) on relative prices for fish, the fishermen could adjust their pattern of production (fishing) and distribution (the market they visit) to secure the highest returns. Mobile phones allowed the fishermen to become very effective rent-seekers, and their rent-seeking activities changed how Kerala's fish markets worked.

A study of 15 beach markets along 225km of the northern Kerala coast found that, once the fishermen used mobile phones, differences in daily prices among the beach markets were cut to a quarter of their previous levels. No boats jettisoned their catches. Reduced waste and the elimination of the dealers' bargaining power raised the profits of fishermen by 8% at the same time as consumer prices fell by 4%.

The mobile phone came close to implementing the Law of One Price in Kerala's fish markets, virtually eliminating the periodic excess demand and supply, to the benefit of fishermen and consumers; but not the fish dealers who had acted as middlemen.

When the Kerala sardine fishermen got mobile phones, they were able to receive the information in the prices in the different beach markets and to respond. The economist Friedrich Hayek had been first to explain this phenomenon: that prices can be *messages*.

GREAT ECONOMISTS

FRIEDRICH HAYEK

The Great Depression of the 1930s ravaged the capitalist economies of Europe and North America, throwing a quarter of the workforce out of work in the US. During the same period the centrally planned economy of the Soviet Union continued to grow rapidly under a succession of five-year plans. Even the arch-opponent of socialism, Joseph Schumpeter, had conceded: "Can socialism work? Of course it can... There is nothing wrong with the pure theory of socialism."

Friedrich Hayek (1899-1992) did not think so. Born in Vienna, he was an Austrian (later British) economist and philosopher who believed that the government should play a minimal role in the running of society. He was against any efforts to redistribute income in the name of social justice. He was also an opponent of the policies advocated by John Maynard Keynes designed to moderate the instability of the economy and the insecurity of employment.

Hayek's book *The Road to Serfdom* was written against the backdrop of the second world war, where economic planning was being used both by German and Japanese fascist governments, by the Soviet communist authorities, and by the British and American governments. He argued that well-intentioned planning would inevitably lead to a totalitarian outcome.

His key idea about economics revolutionised how economists think about markets. It was that *prices are messages*: they convey valuable information about how scarce a good is, information that is available only if prices are free to be determined by supply and demand, rather than by the decision of a planner. Hayek even wrote a comic book, which was distributed by General Motors, to explain how this mechanism was superior to planning.

But Hayek did not think much of the theory of competitive equilibrium that we explained in Unit 8, in which all buyers and sellers are price-takers. "The modern theory of competitive equilibrium, " he wrote, "*assumes* the situation to exist which a true explanation ought to account for as the effect of the competitive process."

In Hayek's view, assuming a state of equilibrium (as Walras had done to create general equilibrium theory) prevents us from analysing competition seriously. He defined competition as "the action of endeavouring to gain what another endeavours to gain at the same time."

Hayek explained:

"Now, how many of the devices adopted in ordinary life to that end would still be open to a seller in a market in which so-called 'perfect competition' prevails? I believe that the answer is exactly none. Advertising, undercutting, and improving ('differentiating') the goods or services produced are all excluded by definition—'perfect' competition means indeed the absence of all competitive activities."
— Friedrich A. Hayek, *The Meaning of Competition* (1946)

The advantage of capitalism, to Hayek, is that it provides the right information to the right people. In 1945 he wrote:

"Which of these systems [central planning or competition] is likely to be more efficient depends on… which of them we can expect [to make] fuller use of the existing knowledge. And this, in turn, depends on whether we are more likely to succeed in putting at the disposal of a single central authority all the knowledge which ought to be used but which is initially dispersed among many different individuals, or in conveying to the individuals such additional information as they need in order to enable them to fit their plans in with those of others."
— Friedrich A. Hayek, *The Use of Knowledge in Society* (1945)

Hayek's challenging ideas, and their application, are still causing arguments today.

In Unit 8 we introduced the concept of market equilibrium: a situation in which the actions of the buyers and sellers of a good have no tendency to change its price, or the quantity traded. We also studied how changes from the outside called *exogenous shocks*—like an increase in the demand for bread or a new tax—will change the equilibrium price and quantity.

The opposite of exogenous is *endogenous* meaning "coming from the inside" and resulting from the workings of the model itself. Here we will study how these endogenous responses to exogenous shocks take place, through the real-world competition that Hayek complained was absent from the model of competitive equilibrium.

9.1 HOW PEOPLE CHANGING PRICES CAN LEAD TO A MARKET EQUILIBRIUM

When Lincoln's decision to blockade the southern ports led to a drastic shortage of cotton on the world market (Unit 8), people saw the opportunity to benefit by changing the price. In turn, these price changes sent a message to producers and consumers around the world to change their behaviour.

We can extend the model of the equilibrium of competitive price-takers from Unit 8. When there is a shift in supply or demand, the people who recognise there are economic rents to be gained can benefit by changing the price. Compare this with an equilibrium in which no one can benefit by offering or charging a different price, given the price everyone else was offering or charging. The market-clearing and price-taking outcome we described in that case was a Nash equilibrium.

We now ask what happens when a market in equilibrium (such as is illustrated by point *A* in Figure 9.2a) experiences an exogenous shock that shifts the demand or supply curve. From the model in Unit 8, we know that a shift in the demand for hats implies there is a new (Nash) equilibrium at a higher price and quantity of hats sold (point *C* in Figure 9.2a).

Work through the slideline to see how the market adjusts to the shock.

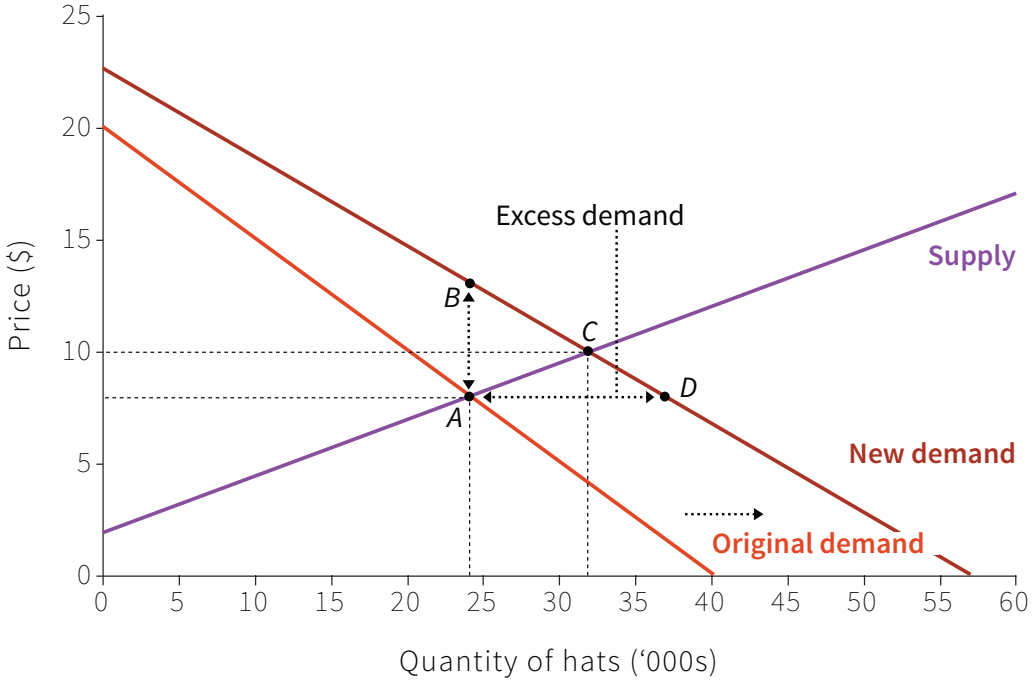


Figure 9.2a *An increase in demand in a competitive market: opportunities for rent-seeking.*

But how do we get from point *A* to point *C* in a decentralised market economy where there are many buyers and many sellers? Neither the hat buyers nor those selling the hats know the new equilibrium price is $10. Moreover at the old competitive equilibrium (*A*) we know from the previous unit that all buyers and sellers were acting as price-takers, that is they were considering the price of all others in the market as given and were in a situation such that they could not benefit from changing their own price. If everyone were to remain a price-taker, the price would not change.

But following the shift in demand at the going price the number of hats demanded exceeds the number supplied (see point *D*)—termed *excess demand*—and *A* is no longer an equilibrium. As a result of the shift in demand:

- Price-*taking* is no longer a Nash equilibrium.
- Some of the buyers or sellers will realise that they can benefit by being a price-*maker*, and decide to charge a different price from the others.

Here is an example of how this might happen.

A hat-seller notices that every day there are customers wishing to buy hats, but there are none left on the shelf. Some of them, she realises, would have been happy to pay more than the going price. And moreover some who paid the going price for their hat would have been willing to pay more too. So the hat-seller will raise her price the next day.

In Figure 9.2a, when demand has increased, a hat-seller who observes more customers can work out that she can make higher profits by raising the price (see point *B* where price is above marginal cost, shown by the supply curve at point *A*). Of course she does not know exactly where the new demand curve is, but she cannot fail to see the people who want to buy hats who go home disappointed.

If she raises the price she would raise her profit rate, and at least temporarily earn an economic rent—that is, make profits greater than the profit necessary to keep her hat business going. Moreover because her price now exceeds her marginal cost she will seek to produce and sell more hats. The same is true of other hat-sellers who will experiment with higher prices and increased outputs.

As a result of the rent-seeking behaviour of hat-sellers, the hat industry adjusts and a new equilibrium will emerge with the price and quantity combination (at point *C* in Figure 9.2a). Once the industry is at point *C*, the market again clears, supply is equal to demand, and at the new equilibrium price none of the sellers or buyers can benefit from charging a price different from $10. They all return to being price-takers, until the next change in supply or demand comes along.

From Figure 9.2b, notice that sellers do not respond to the increase in demand by selling more hats at the existing price, for example, at point *D*. The reason is clear: at point *D*, sellers are making a loss because price is below marginal cost and there will be no willing suppliers.

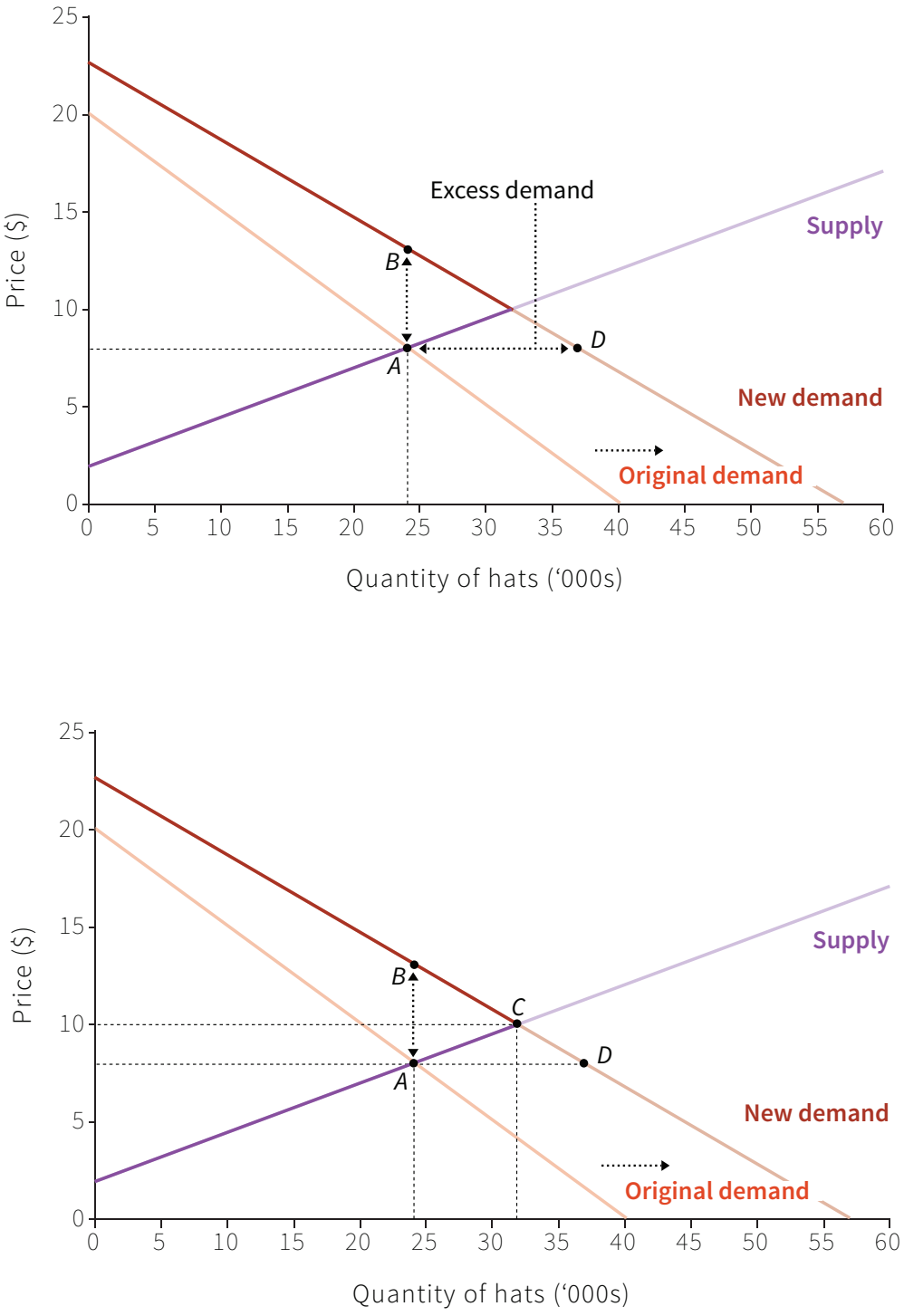


Figure 9.2b *Voluntary trades will occur only on the short side of the market.*

What we know from Figure 9.2b is that, until the new equilibrium (point C) is reached, the price can be anywhere between $8.00 and $13.20, and the amount bought and sold cannot be greater than the quantities indicated by the thick portions of the supply and demand curves in Figure 9.2b. The other parts of the demand and supply curves are irrelevant because if (at a given price) supply exceeds demand, then the amount of hats sold will be determined by those demanding them. And if the reverse is true, then the hat sellers will determine the number of hats sold.

The term the short side of the market is used to refer to the thick portions of the lines in Figure 9.2b, meaning the side (either supply or demand) on which the number of desired transactions is less. You can see from the figure that:

- *If the price is high*: Demanders will be on the short side of the market.
- *If the price is low*: Suppliers will be the short-siders.

If we take the opposite case in which there is a fall in demand for hats, the situation is shown in Figure 9.3. This time, hat-sellers experience a fall in sales. Work through the slideline to see the opportunities that arise for buyers or sellers to benefit by changing the price when there is excess supply in the hat market.

Figure 9.3 *A decrease in the demand for hats.*

In the new situation, acting as a price-taker and charging $8 is no longer a Nash equilibrium; and neither is buying a hat for $8. A customer at the hat shop might say to the hat-seller: "I see you have quite a few unsold hats piling up on your shelf. I'd be happy to buy one of those for $7."

To the buyer this would be a bargain. But it's also a good deal for the seller, because at the reduced level of sales $7 is still greater than the hat seller's marginal cost of producing the hats. So both buyer and seller act as price-makers, transacting at a price different from the previous equilibrium price that the other sellers—acting as price-takers—are charging.

To summarise:

- If a market is in competitive equilibrium, an exogenous change in supply or demand will result in either *excess supply or excess demand* at the previous equilibrium price.
- If there is excess demand, for example, there will be buyers *willing to pay more* for an additional unit of the good than the marginal cost of producing it.
- That difference—the excess of the willingness to pay over the marginal cost—represents a *potential rent* that is an opportunity for gain that is better than the next best alternative, which is continuing to transact at the going price.
- Under conditions of market disequilibrium, *some buyers, or sellers, or both will find that they can increase their profits or their utility* (at least temporarily) by capturing these potential rents.
- They do this *by becoming a price-maker*, charging or offering prices different from the previous equilibrium price.
- This process will go on as long as there is either excess demand or supply; that is, *until a new competitive equilibrium is attained*.

We call this process *market equilibration through rent-seeking*, and the rents that are the moving force in this story are termed *disequilibrium economic rents*.

Economists have studied this process using experiments in which the participants acted as buyers and sellers engaged in rent-seeking when the experimental market was out of equilibrium. See the box *Equilibration through rent-seeking in an experimental market*.

Notice how market equilibration through rent-seeking resembles the process of technological improvement through rent-seeking that we modelled in Unit 2. There the exogenous change was the possibility of introducing a new technology. The first to install the new machine or adopt the new process gained profits in excess of the normal profit rate, termed *innovation rents*. This process went on until the innovation was widely diffused in the industry and prices had adjusted so that there were no further innovation rents to be had. In a competitive market, disequilibrium rents play the same role in creating incentives for action (changes in price and quantity) as innovation rents play in promoting new processes, or new ways of doing business.

DISCUSS 9.1: A SUPPLY SHOCK AND ADJUSTMENT TO A NEW MARKET EQUILIBRIUM

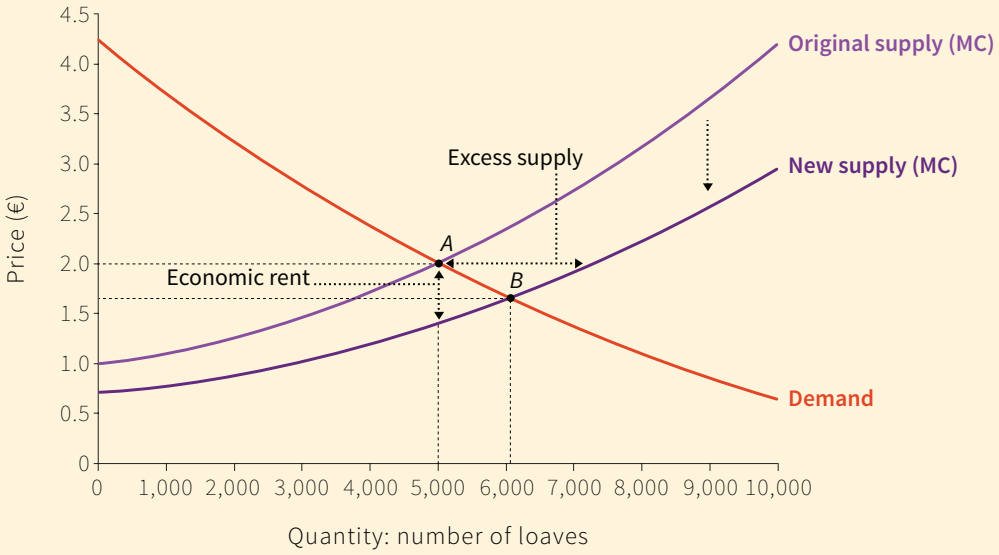


Figure 9.4 *An increase in the supply of bread (fall in MC).*

Consider the case of the bakery industry supplying the restaurant trade, as in Figure 9.4. Suppose there is a new technology available to the bakery industry which results in the supply curve shifting as shown.

1. Explain how the actions of bakeries adjust the industry to a new equilibrium.
2. Is it always the seller who benefits from the economic rents that arise when the market is in disequilibrium?
3. If a restaurant buys bread, what action might it take?

HOW ECONOMISTS LEARN FROM FACTS

EQUILIBRATION THROUGH RENT-SEEKING IN AN EXPERIMENTAL MARKET

Economists have studied the behaviour of buyers and sellers in laboratory experiments to assess whether prices do adjust to equalise supply and demand. In the first such experiment, in 1948, Edward Chamberlin gave each member of a group of Harvard students a card designating them "buyers" or "sellers" and stating their willingness to pay or reservation price in dollars. They could then bargain amongst themselves, and he recorded the trades that took place. He found that prices tended to be lower, and the number of trades higher, than the equilibrium levels. One of the students who took part, Vernon Smith, later conducted his own experiments and won a Nobel prize in economics as a result.

He modified the rules of the game so that participants had more information about what was happening: buyers and sellers called out prices that they were willing to offer or accept. When anyone agreed to a proposed deal, a trade took place and the two participants dropped out of the market. His second modification was to repeat the game several times, with the participants keeping the same card in each round.

Figure 9.5 shows his results. There were 11 sellers, with reservation prices between \$0.75 and \$3.25, and 11 buyers with WTP in the same range. The diagram shows the corresponding supply and demand functions. You can see that, in equilibrium, six trades will take place at a price of \$2. But the participants did not know this, since they did not know the price on anyone else's card. The right-hand-side of the diagram shows the price for each trade that occurred. In the first period there were five trades, all at prices below \$2. But by the fifth period most prices were very close to \$2, and the number of trades was equal to the equilibrium quantity.

While the model of price-taking among competitors fails to capture the rent-seeking activities of the market traders in Smith's experiment, the outcome of their bargaining is correctly predicted by the price-taking equilibrium. This will generally be the case in a market in which goods are identical, there are enough buyers and sellers, and those buyers and sellers are well-informed about the trading of others. The outcome was close to equilibrium even in the first period, and converged quickly towards it subsequently as the participants learned more about supply and demand, just as Marshall argued that it would do.

This model predicts the outcome of competition in markets for many goods and services. But it is far from the general rule. In Unit 8 we saw evidence suggesting that the effects of competition on prices may be weak, and we turn next to some examples where prices do not clear markets.

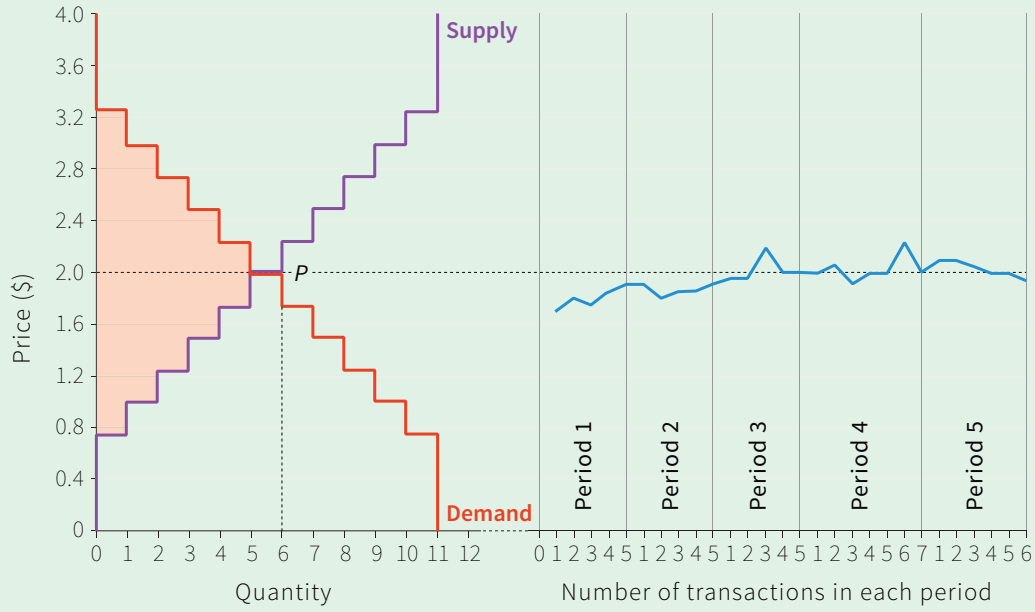


Figure 9.5 *Vernon Smith's experimental results.*

Source: *Smith, Vernon L. 1962. 'An Experimental Study of Competitive Market Behavior.' Journal of Political Economy 70 (3): 322.*

In Smith's experiment, and in the model of the market we have just presented, people continue changing the price and quantities at which they transact until there is neither excess supply nor excess demand, so the market clears. But this is not always the case.

Now we look at two contrasting cases where markets don't clear, but remain in a state of excess supply or excess demand:

- *The market has an equilibrium in which excess supply exists*: Even though the market does not clear, there is no way any buyer or seller can benefit by changing his or her price or quantity.
- *The market does not clear in spite of the fact that there are price or quantity changes that could benefit sellers or buyers*: These adjustments are not legally allowed or do not happen for some other reason. We will consider markets that are prevented from clearing by the intervention of the government or the choice of the firm.

DISCUSS 9.2: COTTON PRICES AND THE AMERICAN CIVIL WAR

Read again the introduction to Unit 8 and use the methods introduced above, with the models from Units 6, 7 and 8, and the box about Hayek, to represent:

1. The increase in the price of US raw cotton (show the market for US raw cotton, a market with many producers and buyers).
2. The increase in the price of Indian cotton (show the market for Indian raw cotton; a market with many producers and buyers).
3. The reduction in textile output in an English textile mill (show a single firm in a competitive product market).

In each case indicate which curve(s) in the relevant figures shifted and explain the result.

9.2 THE LABOUR MARKET: A MARKET THAT DOES NOT CLEAR IN EQUILIBRIUM

In this and the next section we present a model of the labour market in which, in equilibrium, the supply of labour (number of people seeking jobs) exceeds the demand for labour (number of jobs offered). Those without work in this situation are termed the *involuntary unemployed* (to distinguish them from those who are, by choice, temporarily out of work because they left one job to start another one).

Understanding the causes of unemployment in different countries, and at different times, requires us to take into account shocks hitting the economy as a whole—such as the collapse of an investment boom, government and central bank policy, the actions of trade unions, and the introduction of new technologies, as well as the skills and experience of those seeking work. In later units we will show the way these factors can affect unemployment. In this unit, we exclude these factors and study the fundamental determinants of how many job seekers will find jobs, and whether there would be unemployment when the labour market is in equilibrium.

The model of the labour market is quite different from the model of equilibrium of price-taking buyers and sellers in Unit 8. Before we present this new model, think about whether the model of price-taking in competitive equilibrium can work for the labour market.

Recall that *in the equilibrium* of the bread market that we used as an illustration, neither bread consumers nor bakeries selling bread could benefit by offering to pay a different price, or setting a different price from the one that prevailed in other transactions throughout the market. Buyers and sellers were price-takers in equilibrium:

- *No buyer could benefit from asking to pay less than the prevailing price*: No bakery would agree to the sale.
- *No buyer could benefit by offering to pay more than the going price*: This would just be throwing away money. Buyers in the bread market are price-takers because they wish to purchase bread at the lowest possible price.
- *No seller (a bakery) could benefit from setting a higher price*: There would be no customers.
- *No seller could benefit by offering a lower price*: This would be throwing away money. They can have as many customers as they like at the existing price.

Now think about a buyer in the labour market. This is an employer who buys the employee's time. The price is the wage. An employer who acts like a bread buyer would offer the employee the lowest wage that the individual would accept to take the job. This lowest possible wage, recall from Unit 6, is the *reservation wage*.

We know from Unit 6 that an employer who did this would be disappointed. The worker who is paid just a reservation wage does not worry about losing the job, and so would have little incentive to work hard for the employer. Instead, we saw that employers choose a wage to balance their wage costs against the positive effects that a higher wage has on the employee's motivation to work.

In the bread market, the sales contract between buyer and seller is for bread, and if you buy bread you get what you want. It's a *complete* contract. (A contract need not be in writing and it need not be signed to be enforceable: your receipt is enough to get a refund if the bag labelled "fresh bread" turned out to contain a week-old loaf when you got home.)

By contrast, in the labour market, the employment contract is usually for the employee's work time, not for the work itself. Because it is the employee's work that produces the firm's goods and is essential to the firm's profits, this means the contract is *incomplete*: something that matters to one of the parties to the exchange is not covered in the contract.

The implication is that, in contrast to the bread market, for a buyer in the labour market, *paying more than is necessary to buy the employee's time* is not throwing away money; it is the way that employers get what they want—work—and how they make profits. And because they are deciding on the price (that is, the wage) that they will offer the worker, they are not price-takers, they are price-setters. This is why Unit 8's model of the competitive equilibrium of price-takers will not work in the labour market.

9.3 THE LABOUR MARKET MODEL

We model the labour market of an entire economy with price-setting firms, selling differentiated products like those described in Unit 7, and a large number of identical workers who may be employed by the firms for a single wage (as studied in Unit 6).

Two basic concepts are required to understand how the wages (the price) and employment (and quantity) are jointly determined in the labour market:

- The *wage curve*: The wage curve from Unit 6 gives the wage necessary at each level of employment to provide workers with incentives to work hard and well. The wage curve is the model's representation of the relationship between firms and their employees.
- The *profit curve*: The profit curve is the relationship between the wage and the price that results when firms set prices to maximise their profits. As we saw in Unit 7, the weaker the competition faced by a firm, the less elastic its demand, and therefore the higher the price it sets relative to its costs. Those costs include the wage that it pays. The profit curve is the model's representation of the relationship between firms and their customers.

The wage curve—firms and their employees

The wage curve provides the answer to the hypothetical question: "If the level of employment is *this* number of workers (a point on the horizontal axis), then what is the wage that the employers will offer (the corresponding point on the vertical axis) so as to keep the employees working at the least possible cost to the firm?" (Check what it looks like in theory and an empirical estimate in Unit 6).

This is not the same thing as a supply curve for labour, which would be the answer to a different hypothetical question: "If the wage were *this* amount (the point on the vertical axis) how many people would seek work (the corresponding point on the horizontal axis)?"

We will see later that changes in the number of people seeking work—the supply of labour—affect the equilibrium of the labour market, but not directly (as in the bread market). Instead changes in labour supply affect wages and employment indirectly by shifting the wage curve.

The profit curve—firms, costs and customers

But, as a model of the labour market, the wage curve alone is like one hand clapping. To determine the wage and employment level we need another curve. This is the profit curve.

Firms employ workers to produce output, which is sold to make profits. As we saw in Unit 7, the firm sets the price of its product to maximise its profits. This profit-maximising price depends on its costs, and on the elasticity of demand it faces. More precisely, the price is set so that the markup of the price above the marginal cost is inversely proportional to the elasticity. Assume, for simplicity, that the cost of an additional unit of output is the wages of the workers needed to produce it. The firm's choice of price determines a relationship between the wage and the price. *The more competition the firm faces, the lower the price it sets relative to the wage—or, in other words, the higher the real wage, which is the wage relative to the price*.

Just as the wage curve represents the relationship between firms and their employees in the economy as a whole, the profit curve represents the relationship between firms, their costs and their customers. We draw both curves in a diagram with the real wage on the vertical axis and employment on the horizontal axis. Together the two curves determine the outcome—the wage and the level of employment—when the labour market is in equilibrium.

- The profit curve shows the real wage paid when firms choose their profit-maximising price, as shown in Figure 9.6. The corresponding level of profit received by the owners depends on how much competition there is in the economy.

The key ideas for understanding the profit curve are:

- *Competition*: The extent of competition in the economy determines the extent to which firms can charge a price that exceeds their costs, that is, the markup. The less competition, the greater the markup. In Figure 9.6, a higher markup, which results from less competition among firms, will increase the profit per worker. Since this leads to higher prices across the whole economy, it implies lower real wages, pushing down the profit curve.
- *Labour productivity*: For any given markup, the level of labour productivity—how much a worker produces in an hour—determines the real wage. The greater the level of labour productivity, the higher the real wage that is consistent with a given markup. In Figure 9.6, higher labour productivity shifts the dashed line upwards, and, keeping the markup unchanged, the profit curve will shift upwards raising the real wage.

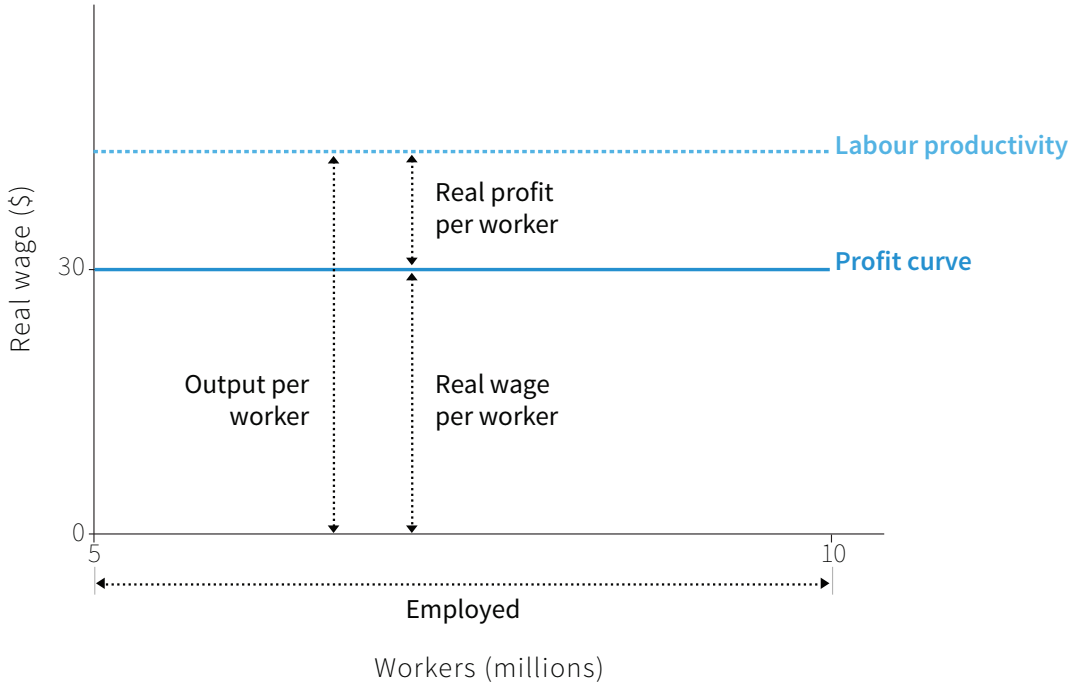


Figure 9.6 *The profit curve.*

Price-setting by firms produces the profit curve. It splits output per worker into real profit per worker, and the real wage per worker.

DISCUSS 9.3: THE PROFIT CURVE

As the two bullet points above explain, the level of competition and labour productivity are held constant in drawing Figure 9.6. As the level of competition or labour productivity changes, the profit curve also changes. This is an example of *ceteris paribus* reasoning: that is, holding other things constant. Make a list of other possible influences on the profit curve that we do not consider here.

To understand more about the profit curve, read the section on it in this unit's Einstein.

9.4 THE LABOUR MARKET: UNEMPLOYMENT AND LABOUR SUPPLY

Labour market equilibrium: The wage curve and the profit curve

Superimposing the wage curve on the profit curve in Figure 9.7 we have a picture of the two sides of the labour market.

Use the slideline to show that when 9 million workers are employed, the real wage that delivers the markup consistent with the extent of competition in the economy (shown by the profit curve) is just high enough to provide workers with the incentive to work and to make hiring them worthwhile for firms (shown by the wage curve).

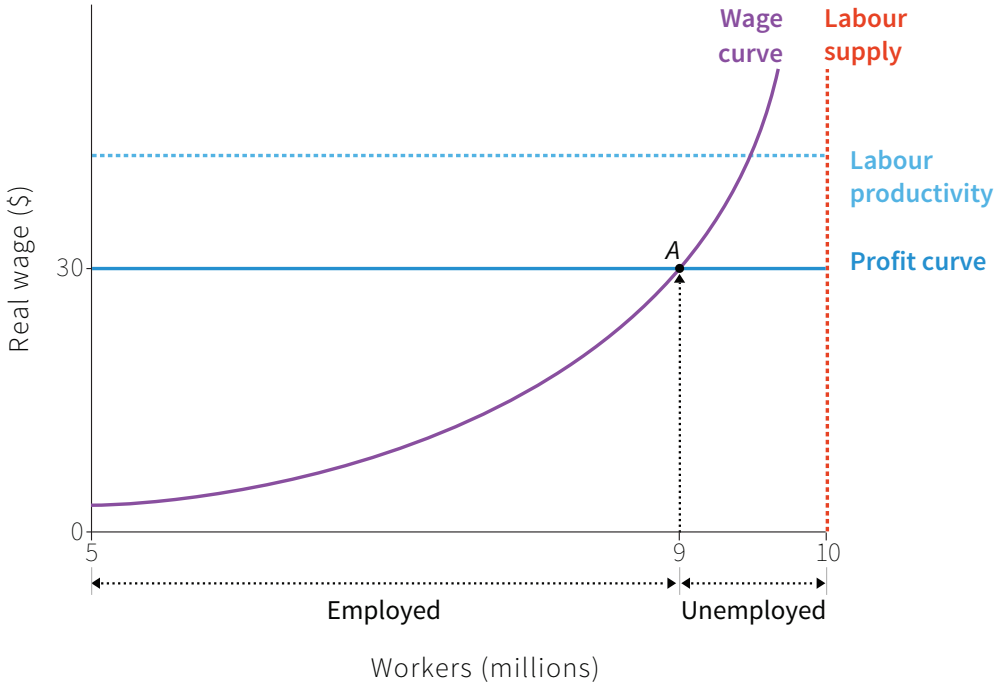


Figure 9.7 *Equilibrium in the labour market.*

At point *A* the real wage that delivers the markup consistent with the extent of competition in the economy is just high enough to provide workers with the incentive to work, and to make hiring them worthwhile for firms.

Follow the slidebar in Figure 9.8 to see what would happen if there were a reduction in the degree of competition faced by firms. The markup would increase, and as a result the profit curve would fall, leading to a new equilibrium at point B with a lower wage and a lower level of employment. The share of output going to profits rises, and the share going to wages falls.

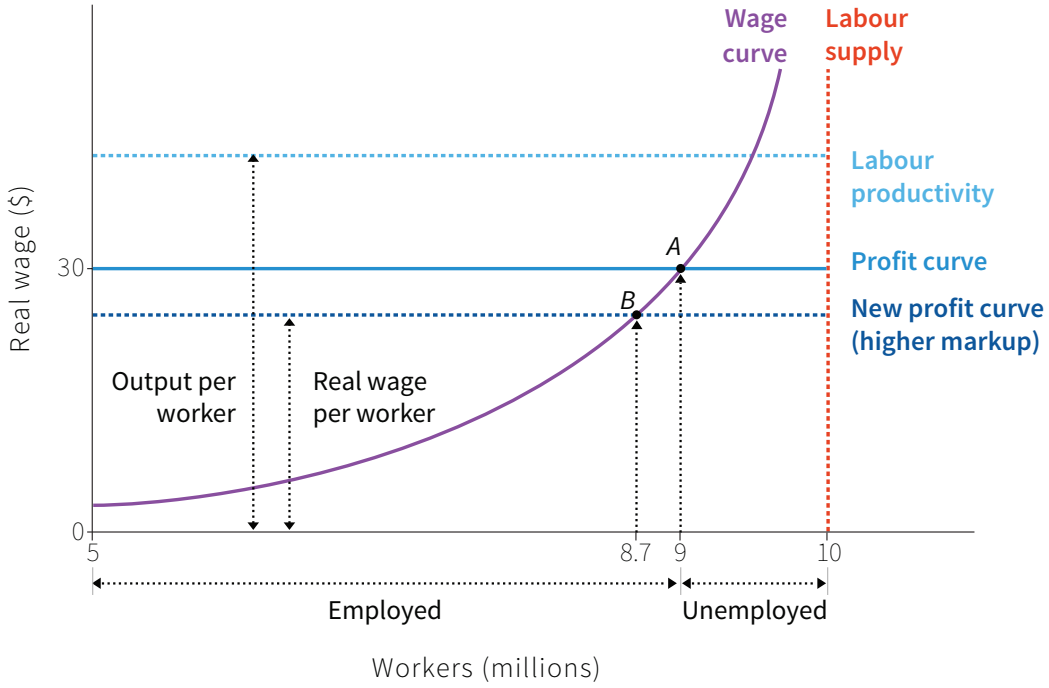


Figure 9.8 *The labour market and its long-run equilibrium*

The intersection of the profit curve and the wage curve is the Nash equilibrium of the labour market because all parties are doing the best they can in the circumstances:

- *The firms are offering the most profitable wage (on the wage curve), given employment*: If 9 million workers have jobs, then the best that firms can do (from the wage curve) is to offer the wage shown by the profit curve.
- *Employment is the highest it can be (on the profit curve), given the wage offered*: Firms will hire 9 million workers.
- *Those who fail to get jobs would rather have a job, but there is no way they can get one*: Not even by offering to work at a lower wage than others.
- *Those who have jobs cannot improve their situation by changing their behaviour*: If they worked less on the job, for example, they would run the risk of becoming one of the unemployed; if they demanded more pay, their employer would refuse or hire someone else.

DISCUSS 9.4: IS THIS REALLY A NASH EQUILIBRIUM?

In this model, the unemployed are no different from the employed (except for their bad luck). Imagine you are an employer, and one of the unemployed comes to you and promises to work at the same effort level as your current workers but at slightly less pay.

1. Would you accept the offer? Explain your answer.
2. Given your answer, can you explain why unemployment exists in a Nash equilibrium?

DISCUSS 9.5: A REDUCTION IN UNEMPLOYMENT BENEFIT

In Unit 6 we showed that a reduction in the unemployment benefit would shift the wage curve downward.

1. How would this affect the wage rate and the level of employment in the labour market? Consider three different kinds of actors in the status quo prior to the reduction in the unemployment benefit—employers, employed workers and the unemployed—and explain how each of these groups might be affected by such a change.
2. How would you expect someone from each of these groups to vote on the proposal to reduce unemployment benefits?

Unemployment as a characteristic of labour market equilibrium

We have shown unemployment can exist in a Nash equilibrium in the labour market. In Unit 6, considering an individual firm, using proof by contradiction we saw that if everybody had a job (no unemployment) then nobody would work hard enough for firms to make profits and therefore nobody would be employed! We now show why there will always be unemployment in labour market equilibrium in the economy as a whole.

Unemployment means that there are people seeking work and not finding it, termed *excess supply* in the labour market. To understand why there will always be unemployment in labour market equilibrium, we introduce the labour supply curve.

In our model it is vertical, meaning that higher wages do not on balance lead more people to offer more hours at work. At higher wages some people seek (and find) more hours of work, and others seek (and find) shorter hours. You know from Unit 3 that the substitution effect of a wage increase (leading to the choice of more hours of work and less of free time) may be offset by the opposite income effect. For simplicity we draw a supply curve such that, on balance, the wage has no effect on the labour supply. But this is not important. The model would not be different if higher wages led to either more or fewer people seeking work.

Here is why, in labour market equilibrium, there will always be some unemployment:

- If there is no unemployment there can be no cost of job loss (no employment rent) because a worker who loses her job immediately gets another one at the same pay.
- Therefore some unemployment is necessary in order for the employer to be able to motivate workers to provide effort on the job.
- This means that the wage curve is always to the left of the labour supply curve.
- Labour market equilibrium is always on the wage curve.
- In any equilibrium, where the wage and profit curves intersect, there are unemployed people, shown by the gap between the wage curve and the labour supply curve.

Why labour supply matters

Even though the demand for labour must fall short of the supply, the labour supply is one of the determinants of the Nash equilibrium of the labour market. To see why this is so, imagine that there is immigration; or that people who have stayed at home to raise children, or have retired, rejoin the labour force. The labour supply curve would shift to the right, as shown in Figure 9.9.

What effect would this have? Let's look first at what happens to the wage curve following an increase in labour supply:

- The new jobseekers would enter the pool of unemployed.
- This would increase the expected duration of a spell of unemployment.
- By raising the cost of job loss, this increases the employment rent enjoyed by employed workers at the current wage and level of employment.
- But firms would then be paying more than necessary to ensure worker motivation on the job...
- ... Therefore firms would lower their wages.

Because this is true of any point on the wage curve it must be true of the entire curve. So the effect of an increase in labour supply is to shift the wage curve downward.

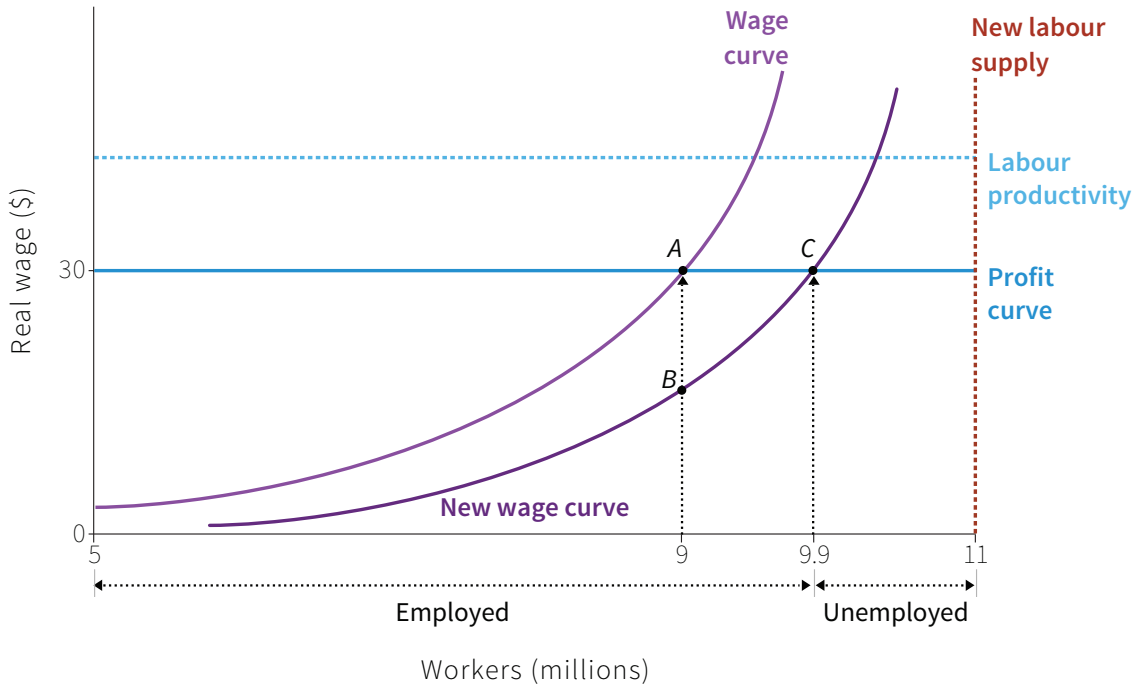


Figure 9.9 *The effect of an increase in labour supply.*

We summarise the effects of the increase in labour supply on the labour market:

- The shift downward in the wage curve at the initial level of employment lowered the wage (to *B*).
- The reduction in the wage results in a reduction in the firms' marginal costs and with no change in the firm's demand conditions, the firms will hire additional workers.
- As a result, employment expands so that once again the economy is at the intersection of the profit curve and the new wage curve, with higher employment.
- The increase in labour supply leads to a new equilibrium at higher employment because it shifts the wage curve down. New hiring stops when the wage is once again at the level shown by the profit curve (at C). In the new equilibrium, employment is higher and the real wage is unchanged.

The labour market is a very important case in which a market does not clear, even when in equilibrium. Notice in our model that:

- Wages were free to adjust to any level that the participants wished to implement.
- Firms were unrestricted in their decisions to hire and fire workers.
- There were no government-imposed minimum wages, or trade unions pushing wages up.

So in the model of the labour market, *unemployment is not the result of government policies, unions or high wages*. It is the unintended but unavoidable result of that fact that firms cannot write complete contracts with their employees. An essential part of

the way profit-maximising firms motivate workers to do a good job is by conferring employment rents on workers. This works because employees have something to lose—their employment rents—if they lose their job. This means that the wage curve is always to the left of the labour supply curve.

We began by asking why the model of the bread market with price-taking buyers and sellers and market clearing in equilibrium does not work for the labour market. Figure 9.10 summarises the differences:

	BREAD MARKET: AN EQUILIBRIUM OF PRICE-TAKERS	MARKET FOR BARISTAS: PRICE-SETTING BY EMPLOYERS AND UNEMPLOYMENT
BUYERS	Individual consumers	Firms (employers)
SELLERS	Firms (shops)	Individual workers
PRICE	Price per loaf	Wage (hourly/weekly/monthly)
WHAT IS BEING SOLD?	A loaf of bread	An hour, week or month of the worker's time
WHAT DOES THE BUYER WANT?	A loaf of bread	The employee's effort on the job
HOW MUCH COMPETITION IS THERE AMONG THE SELLERS?	There are many bakeries competing to sell bread	There are many actual or would-be baristas competing to sell their time
IS THE CONTRACT BETWEEN THE BUYER AND SELLER COMPLETE?	*Yes:* If the bag labelled bread did not contain bread, you get your money back	*No:* The firm's profits depend on the worker's effort per hour/week/month worked, which is not covered in the contract
PRICE-TAKING BUYERS?	*Yes:* Individual bread-buyers cannot bargain for a lower price than others are willing to pay (and would not want to pay more)	*No:* The buyer (the firm) sets the wage to minimise the cost of getting the worker to work; it cannot benefit by offering the lowest wage at which the worker (the seller) would accept the job
IS THERE EXCESS SUPPLY OR DEMAND IN EQUILIBRIUM?	*No:* The market clears. Sales take place at the lowest price the seller would accept	*Yes:* There is excess supply because firms offer a wage higher than the worker's reservation wage (the lowest price the seller would accept) to maximise their profits
CAN WE USE THE COMPETITIVE MODEL FROM UNIT 8?	*Yes:* Not all shops and loaves of bread are identical, but the model provides reasonable predictions of what we observe in bread markets	*No:* The prediction that the market clears is contrary to what we observe when we look at real labour markets in which there is typically substantial excess supply (unemployment)

Figure 9.10 *Two models of market competition: bread and baristas (café or bar workers).*

9.5 OTHER NON-CLEARING MARKETS: RATIONING, QUEUING AND SECONDARY MARKETS

Tickets for the 2013 world tour by Beyoncé sold out in 15 minutes for the Auckland show in New Zealand, in 12 minutes for three UK venues, and in less than a minute for Washington, DC in the US. When American singer Billy Joel announced a surprise concert in his native Long Island, New York in October 2013, all available tickets were snapped up in minutes. In both cases it's safe to say that there were many disappointed buyers who would have paid well above the ticket price: at the price chosen by the concert organisers, demand exceeded supply.

We see excess demand for tickets for sporting events, too. The London organising committee for the 2012 Olympic games received 22 million applications for 7 million tickets. Figure 9.11 is a stylised representation of the situation for one Olympic event. The number of available tickets, 40,000, is fixed by the capacity of the stadium. The ticket price at which supply and demand are equal is £225. The organising committee do not choose this price, but a lower price of £100; at this price 70,000 tickets are demanded. There is excess demand of 30,000 tickets.

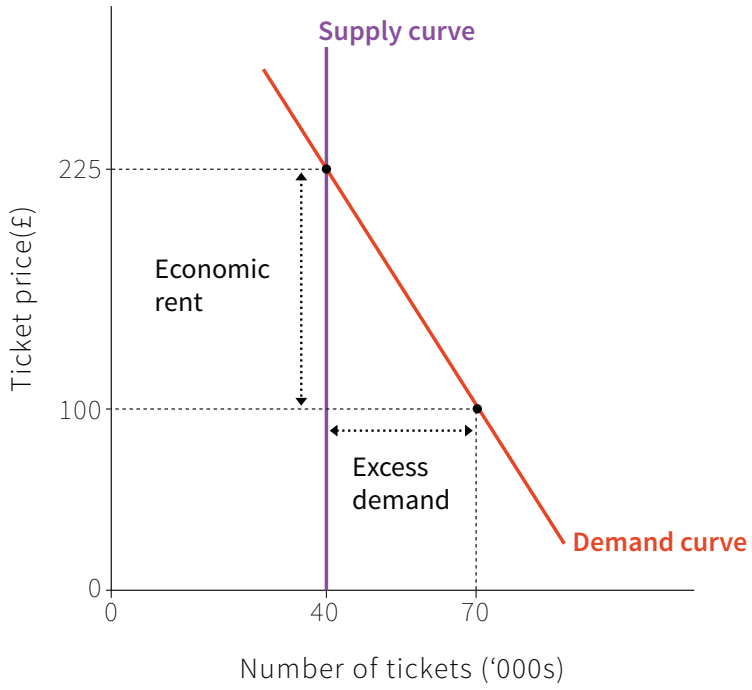


Figure 9.11 *Excess demand for tickets.*

Some of those who succeed in obtaining tickets for a popular event may be tempted to sell them rather than use them. In Figure 9.11, a ticket bought for £100 could be sold for at least £225, making the seller a profit of £125.

The money received by someone behaving like this (the £125) is an example of an economic rent. In this case the next best alternative would be to hold on to the ticket and attend the concert, for which he or she was willing to pay £100. So a person who valued attending the event at £100, and sold a ticket for £225, received an economic rent of £125. The other £100 compensates this person for not seeing the event.

The potential for rents may create a parallel or secondary market. In the case of tickets for concerts and sporting events, part of the initial demand comes from scalpers: people who plan to resell at a profit. Tickets appear almost instantly on peer-to-peer trading platforms such as StubHub or Ticketmaster listed at prices that may be multiples of what was originally paid. In the last few days of the 2014 Winter Olympics in Sochi, tickets for the Olympic Park with a face value of 200 roubles were sold outside the Park for up to 4,000 roubles. (Event organisers may try to prevent this practice; in Sochi the security officers were supposed to intervene.)

Prices in the secondary market equate demand and supply, and allocations are accordingly made to those with the greatest willingness to pay. The assumption that this market-clearing price will be much higher than the listed price is responsible, in part, for the initial frenzied demand for tickets. Nevertheless, some individuals who buy at the lower prices hold on to their tickets, and attend an event that they would otherwise be unable to afford.

In the case of the London Olympics, the organising committee set the price and the tickets were allocated by lottery. This is an example of goods being rationed, rather than allocated by price. The organisers could have chosen a much higher price (£225 for the event in Figure 9.11), which would have cleared the market. But that would have meant that people willing to pay less than £225 would not have seen the event. By allocating the tickets through a lottery some people with a lesser willingness to pay (perhaps because they had limited incomes) would also get to see the Games. There was much public debate about the process, and some anger, but IOC president Jacques Rogge defended it as “open, transparent and fair”.

DISCUSS 9.6: IOC POLICY

1. Do you think the IOC policy of using a lottery is fair?
2. Is it Pareto efficient? Explain why or why not.
3. Using the criteria of fairness and Pareto efficiency, how would you judge the widely criticised practice of “scalping” tickets.
4. Can you think of any other arguments for or against scalping?

There are other cases where the producer of a good chooses to operate with persistent excess demand. The New York restaurant Momofuku Ko offers a 16-course tasting menu at lunch for $175, and has just 12 seats. Online reservations may be made one week in advance, open at 10am daily, and typically sell out in three seconds. In 2008 the proprietor David Chang sold a reservation at a charity auction for $2,780. Even taking into account the willingness of individuals to pay more for an item when the proceeds go to charity, this suggests substantial excess demand for reservations—but he has not raised the price.

DISCUSS 9.7: THE PRICE OF A TICKET

Explain why the seller of a good in fixed supply (such as concert tickets or restaurant reservations) might set a price that the seller knows to be too low, to clear the market.

9.6 MARKETS WITH CONTROLLED PRICES

In December 2013, on a cold and snowy Saturday in New York City, demand for taxi services rose appreciably. The familiar metered yellow and green cabs, which operate at a fixed rate (subject to minor adjustments for peak and night-time hours), were hard to find. Those looking for taxis were accordingly rationed, or faced long waiting times.

But there was an alternative available—another example of a secondary market: an on-demand, app-based taxi service called Uber, which in August 2015 operated in 59 countries. This recent entrant in the local transportation market uses a secret algorithm that responds rapidly to changing demand and supply conditions. Standard cab fares do not change with the weather, but Uber's prices can change substantially. On this December night Uber's surge-pricing algorithm resulted in fares that were more than eight times the base rate charged by the yellow and green cabs. This spike in pricing choked off some demand and also led to some increased supply, as drivers who would have clocked off remained on the road and were joined by others.

DISCUSS 9.8: WHY NOT RAISE THE PRICE?

Discuss the following: "The sharp increase in cab fares on a snowy day in New York led to severe criticism of Uber on social media, but a sharp increase in the price of gold has no such effect."

City authorities often regulate taxi fares as part of their transport policy, for example to maintain safety standards, and minimise congestion. In some countries local or national government also controls housing rents. Sometimes this is to protect tenants, who may have little bargaining power in their relationships with landlords, or sometimes because urban rents would be too high for key groups of workers.

Figure 9.12 shows a situation in which local government might decide to control the housing rent in a city. Note that here we mean rent in the everyday sense of a payment from tenant to landlord for use of the accommodation. Initially the market is in equilibrium at *A*, with 8,000 tenancies at rent of €500—the market clears. Now suppose that there is an increase in demand for tenancies. The supply of housing for rent is inelastic, at least in the short run: since it would take time to build new houses, the only way that more can be supplied in the short run is if some owner-occupiers decide to become landlords and live elsewhere themselves. So the new market clearing rent, €830, is much higher (at *B*).

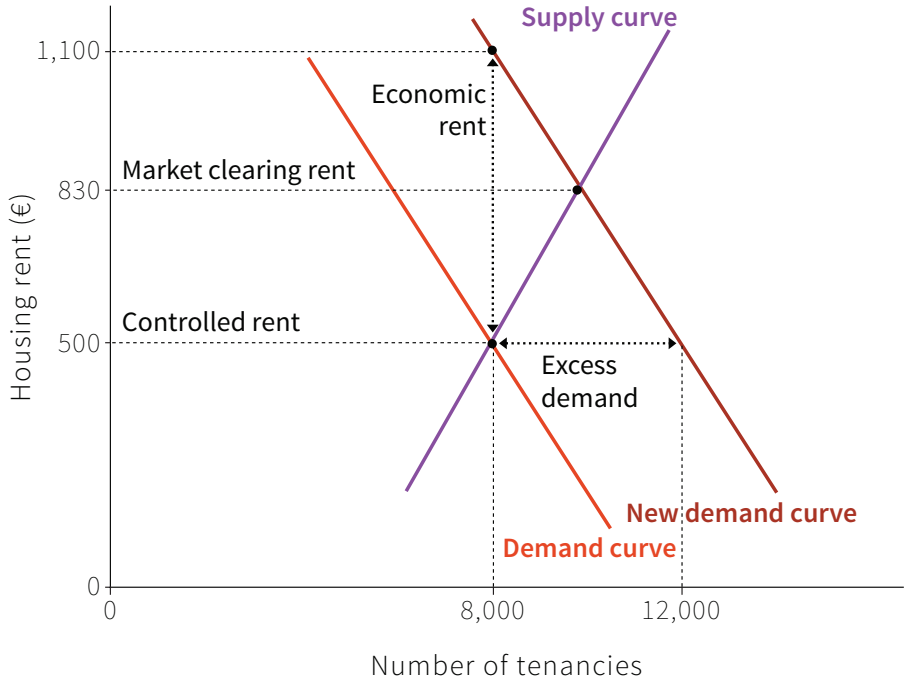


Figure 9.12 *Housing rents and economic rents.*

Suppose that the city authorities are concerned that this rise would be unaffordable for many families, so they impose a rent ceiling at €500. The controlled price of €500 is below the market-clearing price of €830; the supply of housing for rent remains at 8,000, and there is excess demand for tenancies.

In this situation tenancies are not allocated to those with highest willingness to pay. The supply of tenancies is effectively restricted to 8,000 by rent control, and there are 8,000 people willing to pay more than €1,100. But the 8,000 people lucky enough to obtain tenancies may be anywhere on the new demand curve above €500. But unlike the Olympic tickets, rentals are not normally allocated using a lottery.

The rent control policy puts more weight on maintaining a rent that is seen to be fair, and affordable by existing tenants who might otherwise be forced to move out, than it does on Pareto efficiency. The scarcity of rental accommodation gives rise to a potential economic rent: if it were legal (which it usually isn't) some tenants could sublet their accommodation, obtaining an economic rent of €600 (the difference between €1,100 and €500.

If the increase in demand proves to be permanent, the long-run solution for the city authorities may be policies that encourage house-building, shifting out the supply curve so that more tenancies are available at a reasonable rent.

9.7 MARKETS FOR FINANCIAL ASSETS: CHANGING SUPPLY AND DEMAND

Prices in some markets are constantly changing. The graph in Figure 9.13 shows how News Corp's share price on the Nasdaq stock exchange fluctuated over one day in May 2014 and, in the lower panel, the number of shares traded at each point. Soon after the market opened at 9.30am the price was $16.66 per share. As investors bought and sold shares through the day, the price reached a low point of $16.45 at both 10am and 2pm. By the time the market closed, with the share price at $16.54, nearly 556,000 shares had been traded.

Remember from Unit 6 that owning a share in a firm, also known as common stock, means that the shareholder owns a fraction of the firm's capital stock and also gives the shareholder a right to receive a dividend (we use the term *share* and *stock* interchangeably). A portion of the firm's profits, its earnings after payment of interest and taxes, is paid out to shareholders as dividends, while the rest is reinvested in the firm to maintain and expand its ability to generate future profits.

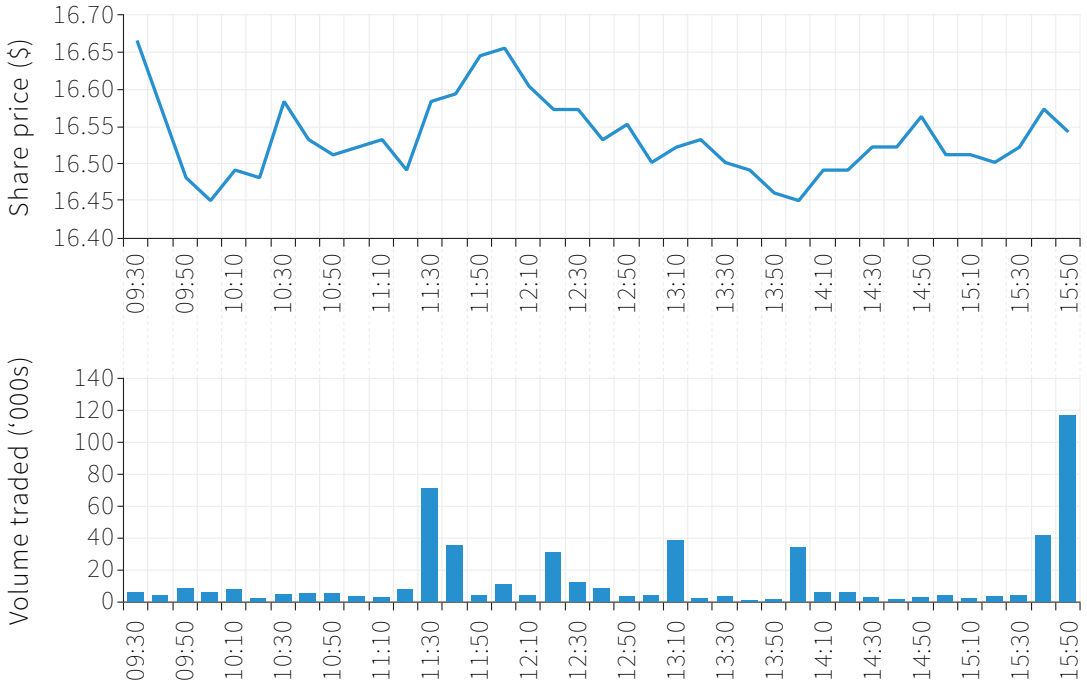


Figure 9.13 *News Corp's share price and volume traded, 7 May 2014.*

Source: *Bloomberg L.P., accessed 28 May 2014.*

The price at which shareholders are willing to buy or sell depends on two things:

- *What they believe about the company's future profitability*: This will influence how valuable their stock is should they choose to sell it, and also the dividends they will get.
- *Their beliefs about how the price might change independently of the firm's profitability*: This depends on what other people believe about future profits.

Think how different this is from the market in bread. You purchase a loaf of bread because you anticipate enjoying eating it, something about which you have little doubt—not because you think the price of bread might rise and you could resell it at a profit tomorrow. Uncertainty about future profitability, and the role played by the beliefs of other market participants about how they will value an asset, distinguishes markets for financial assets. It also distinguishes durable assets (houses or paintings) from markets for goods and services.

To understand how prices of financial assets change, two terms are important:

- *Fundamental value of the shares in a firm*: This is what a well-informed observer would be willing to pay, given what are called the *economic fundamentals* of the firm. Economic fundamentals are current and likely future conditions affecting its profitability including the costs of its inputs, the demand for its outputs, how well run it is, and the prospects for cost-reducing innovations in the future.

- *Speculation*: You would be willing to pay more than the fundamental value of a share if you believe that the share price would rise further above the fundamental value. In this case a share buyer could make a gain by buying at a low price and selling at a higher price, even if the fundamental value of the shares had not changed. Speculation is buying an asset not for use, but to resell in the expectation of a price increase, or selling in expectation of a price decrease.

At any time when the market for shares in News Corp is open, each of the existing shareholders has a reservation price, namely the least price at which the shareholder would be willing to sell. Others are in the market to buy, as long as they can find an acceptable price. Follow the slideline in Figure 9.14 to see the way the price is affected by shifts in demand and supply. The curves show the hourly volume of shares that would be demanded and supplied at each price.

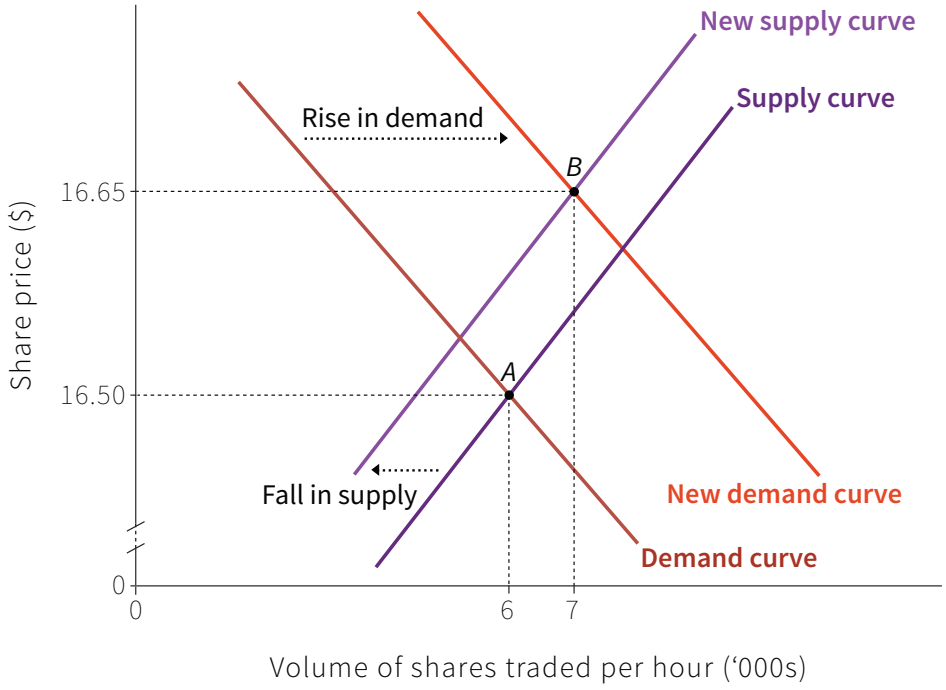


Figure 9.14 *Good news about profitability.*

Initially the market is in equilibrium at *A*: 6,000 shares are sold per hour, at a price of $16.50. Some good news about the future profitability of News Corp simultaneously shifts the demand curve...and the supply curve. The new temporary market equilibrium is at *B*; the price has risen from $16.50 to $16.65. In this illustration demand changes more than supply, so volume rises too.

In practice stock markets don't operate in fixed time periods, such as an hour. Trade takes place continuously and prices are always changing, as we saw in Figure 9.13. To understand how prices change we need to understand a trading mechanism known as a continuous double auction.

This is how the process works (if you prefer, watch this video in which Rajiv Sethi, one of the authors of this unit, demonstrates how orders are processed in a continuous double auction). Anyone wishing to buy can submit a price and quantity combination known as a *limit order*. For instance, a limit order to buy 100 shares in News Corp (NWS) at a price of $16.50 per share indicates that the buyer commits to buying 100 shares, as long as they can be obtained at a price no greater than $16.50 per share. This is the buyer's reservation price. Similarly, a limit sell order indicates a commitment to sell a given quantity of shares, as long as the price is no less than the amount specified. This is the seller's reservation price.

When a limit buy order is placed, one of two things can happen. If a previously placed limit sell order exists that has not yet been filled, and it offers the required number of shares at a price that is at (or below) the amount indicated by the buyer, a trade occurs. If there is no such order available, then the limit order is placed in what is called an *order book* (which is really just an electronic entry), and becomes available to trade against new sell orders that arrive.

Orders to buy are referred to as *bids*, and orders to sell as *asks*. The order book lists bids in decreasing order of price, and asks in increasing order. The top of the book for shares in NWS at around midday on 8 May 2014 looked like Figure 9.15:

Bid		Ask	
PRICE($)	SIZE	PRICE($)	SIZE
16.56	400	16.59	500
16.55	400	16.60	700
16.54	400	16.61	800
16.53	600	16.62	500
16.52	200	16.63	500

Figure 9.15 *A continuous double auction order book: Bid and ask prices for News Corp (NWS) shares.*

Source: *Yahoo! Finance, accessed 8 May 2014.*

Given this situation, a buy order for 100 shares at $16.57 would remain unfilled and would enter the book at the top of the bid column. However, a bid for 600 shares at $16.60 would be filled immediately, since it can be matched against existing limit sell orders. 500 shares would trade at $16.59 apiece, and 100 shares would trade at $16.60. Whenever a buy order is immediately filled, trade occurs at the best possible price for the buyer—the ask price; similarly if a sell order is placed and immediately filled from existing orders, trade occurs at the best possible price for the seller—the bid price. Just as in the market for hats in section 9.1, trades take place on the short side of the market.

We can now see how prices in such a market change over time. If someone receives negative news about News Corp, for example a rumour that an important member of the board is about to resign, and believes that this information has not yet been incorporated into the price, that person may place a large sell order at a price below $16.56, which will immediately trade against existing bids. As these trades occur, bids are removed from the order book and the price of the stock declines.

Similarly, in response to good news, orders to buy at prices above the lowest ask will trade against existing sell orders, and transactions will occur at successively increasing prices.

Since the price is fluctuating, it is not easy to think of this market as being in equilibrium. But it is nevertheless the case that the price is always adjusting to reconcile supply and demand and hence clear the market. Note that the order book (as in Figure 9.13) does not incorporate all of the potential supply of and demand for shares. For example, some actors will not place limit orders because they do not wish to reveal their hand.

Financial assets provide another example of markets equilibrating through economic rent-seeking:

- Those who believe they will benefit from *buying* News Corp shares at a particular price lodge a *bid* at that price.
- Those who believe they will benefit by *selling* lodge an *ask* at a particular price.

The price in the market at any moment in time reflects the aggregate outcome of the rent-seeking behaviour of all the actors in the market—including those who are simply holding on to their shares.

DISCUSS 9.9: SUPPLY AND DEMAND CURVES

1. Use the data from the NWS order book in Figure 9.15 to plot supply and demand curves for shares.
2. Explain why the two curves do not intersect.

9.8 MARKETS FOR ASSETS: HOW PRICE BUBBLES CAN OCCUR

The example of shares in News Corp demonstrates the flexibility of stock prices. This flexibility is common in markets for other financial assets such as government bonds, currencies under floating exchange rates, commodities such as gold, crude oil and corn, and tangible assets such as houses and works of art.

But the prices of shares are not only volatile hour-by-hour and day-by-day. They can also display large swings, often referred to as *bubbles*. To get a sense of the extent of volatility in asset prices, consider Figure 9.16, which shows the value of the Nasdaq Composite Index between 1995 and 2004. This index is an average of prices for a set of stocks, with companies weighted in proportion to their market capitalisation. The Nasdaq Composite Index at this time included many fast-growing and hard-to-value companies in technology sectors.

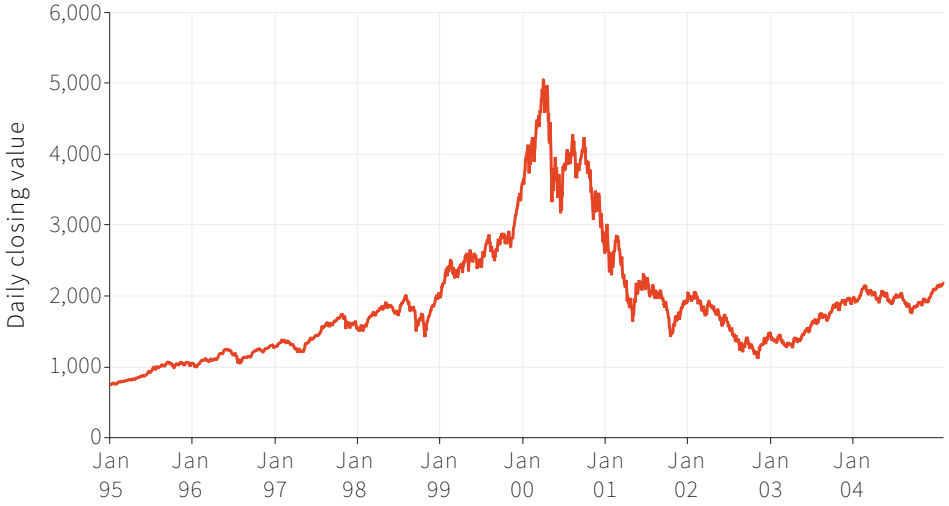


Figure 9.16 *The tech bubble: Nasdaq Composite Index (1995-2004).*

Source: *Yahoo! Finance, accessed 14 January 2014.*

The index began the period at less than 750, and had risen in five years to more than 5,000. The index increased more than six-fold between 1995 and 2000 with a remarkable annualised rate of return of around 45%. It then lost two-thirds of its value in less than a year, and eventually bottomed out at around 1,100, almost 80% below its peak. The episode has come to be called the *tech bubble*.

Information, uncertainty and beliefs

Sometimes new information about the fundamental value of an asset is quickly and reliably expressed in markets. Changes in beliefs about a firm's future earnings growth result in virtually instantaneous adjustments in its share price. Both good and bad news about patents or lawsuits, the illness or departure of important personnel, earnings surprises, or mergers and acquisitions can all result in active trading—and swift price movements.

Because price movements of stocks often reflect important information about the financial health of a firm, traders who do not have this information can try to deduce it from price movements. Using Hayek's language, they are using changes in prices as *messages* from which they can get information. If markets are to work well, traders must respond to these messages. But, when the traders in a market interpret a price increase as a sign of future price increases (called *momentum trading* strategies) the result can be self-reinforcing cycles of price increases, resulting in asset price bubbles and the sudden price declines that typically follow bubbles, called crashes.

Therefore the term *bubble* refers to a sustained and significant departure of an asset price from its fundamental value (the term does not apply only to the stock market).

There are three distinctive and related features of markets which may give rise to bubbles:

- *Resale value*: The demand for the asset arises both from *the benefit* to its owner (for example the flow of dividends from a stock, or the enjoyment of having a painting by a well known artist in your living room) and because it offers the *opportunity for speculation* on a change in its price.

  This applies to more than stocks and Picasso paintings. For example, a landlord may buy a house both for the rental income from it and, in the belief that the price of the house will rise, also to create a capital gain by holding the asset for a period of time and then selling it. People's beliefs about what will happen to asset prices will differ. Also one person's beliefs can change as that person receives new information or believes others are responding to new information.
- *Ease of trading*: In the case of financial assets, the ease of trading means that you can switch between being a buyer and being a seller if you change your mind about whether you think the price will rise or fall. Switching from being a buyer to being a seller (or vice versa) is not possible in markets for ordinary goods and services, where firms with specialised capital goods, and workers with specific skills, are the suppliers who sell to other firms and to households on the demand side.

- *Ease of borrowing to finance sales*: The possibility of borrowing in order to buy assets increases the likelihood of bubbles occurring in these markets. If market participants are able to borrow and increase their demand for an asset that they believe will increase in price, this allows an upward movement of prices to continue, creating the possibility of a bubble and a subsequent crash.

WHEN ECONOMISTS DISAGREE

DO BUBBLES EXIST?

Looking at the price movements in Figure 9.16 (and Figure 9.20, below), one gets the impression that asset prices can be subject to wild swings that bear little relation to the stream of income that might reasonably be expected from holding them.

But do bubbles really exist, or are they an illusion based only on hindsight? In other words, is it possible to know that a market is experiencing a bubble *before* it crashes? Perhaps surprisingly, some of the most prominent economists working with financial market data disagree on this question. They include Eugene Fama and Robert Shiller, two of the three recipients of the 2013 Nobel prize.

Fama denies that the term "bubble" has any useful meaning at all:

"These words have become popular. I don't think they have any meaning… It's easy to say prices went down, it must have been a bubble, after the fact. I think most bubbles are twenty-twenty hindsight. Now after the fact you always find people who said before the fact that prices are too high. People are always saying that prices are too high. When they turn out to be right, we anoint them. When they turn out to be wrong, we ignore them. They are typically right and wrong about half the time."
— Eugene Fama, quoted in 'Interview with Eugene Fama', *The New Yorker* (2010)

This is an expression of what economists call the *efficient market hypothesis*, which claims that all generally available information about fundamental values is incorporated into prices virtually instantaneously. Robert Lucas—another Nobel laureate who is firmly in Fama's camp—explained the logic of this argument in 2009, in the middle of the financial crisis:

"One thing we are not going to have, now or ever, is a set of models that forecasts sudden falls in the value of financial assets, like the declines that followed the failure of Lehman Brothers in September. This is nothing new. It has been known for more than 40 years and is one of the main implications of Eugene Fama's efficient-market hypothesis… If an economist had a formula that could reliably forecast crises a week in advance, say, then that formula would become part of generally available information and prices would fall a week earlier."
— Robert Lucas, *In Defence of the Dismal Science* (2009) Markus Brunnermeier, Mind the frictions (2009)

In a reply to Lucas, Markus Brunnermeier explains why the logic of this position is not watertight:

"Of course, as Bob Lucas points out, when it is commonly known among all investors that a bubble will burst next week, then they will prick it already today. However, in practice each individual investor does not know when other investors will start trading against the bubble. This uncertainty makes each individual investor nervous about whether he can be out of (or short) the market sufficiently long until the bubble finally bursts. Consequently, each investor is reluctant to lean against the wind. Indeed, investors may in fact prefer to ride a bubble for a long time such that price corrections only occur after a long delay, and often abruptly. Empirical research on stock price predictability supports this view. Furthermore, since funding frictions limit arbitrage activity, the fact that you can't make money does not imply that the 'price is right'. "This way of thinking suggests a radically different approach for the future financial architecture. Central banks and financial regulators have to be vigilant and look out for bubbles, and should help investors to synchronise their effort to lean against asset price bubbles. As the current episode has shown, it is not sufficient to clean up after the bubble bursts, but essential to lean against the formation of the bubble in the first place."
— Markus Brunnermeier, *Mind the frictions* (2009)

Shiller has argued that relatively simple and publicly observable statistics, such as the ratio of stock prices to earnings per share, can be used to identify bubbles as they form. Leaning against the wind by buying assets that are cheap based on this criterion, and selling those that are dear, can result in losses in the short run but with long-term gains that, in Shiller's view, exceed the returns that one would make by simply investing in a diversified basket of securities with similar risk attributes.

In collaboration with Barclays Bank, Shiller has launched a product called an exchange traded note (ETN) that can be used to invest in accordance with his theory. This asset is linked to the value of the cyclically adjusted price-to-earnings (CAPE) ratio, which Shiller believes is predictive of future prices over long periods. So this is one economist who has put his money where his mouth is: you can follow the fluctuation of Shiller's index here, and read more arguments for and against the existence of bubbles here.

So there are two quite different interpretations of the data shown in Figure 9.16, about the episode referred to by some as the tech bubble:

- *Fama's view*: Asset prices throughout the tech bubble were based on the best information available at the time and fluctuated because information about the prospects of the companies was changing sharply. In John Cassidy's 2010 interview with Fama in The New Yorker he describes many of the arguments in favour of the existence of bubbles as "entirely sloppy".

- *Shiller's view*: Prices in the late 1990s had been driven up simply by expectations that the price would rise farther still. He called this "irrational exuberance" among investors: the first chapter of the book he wrote using this phrase as its title explains the idea.

Does the price-taking model work for financial and other durable asset markets?

To see whether the model (Unit 8) of price-taking buyers and sellers works for financial and other durable asset markets, look at Figure 9.17. Initially the price of a share in a (so far) hypothetical firm called the *Flying Car Corporation* (FCC) is $50 on the lowest demand curve. When potential traders and investors receive good news about expected future profitability, the demand curve shifts to the right, and the price increases to $60. (For simplicity we assume that the supply curve doesn't move). The rise in price to $60 in response to an exogenous rise in demand is just like the market for bread analysed in section 9.1. Use the slideline to see what happens next –and how this differs from the market for bread.

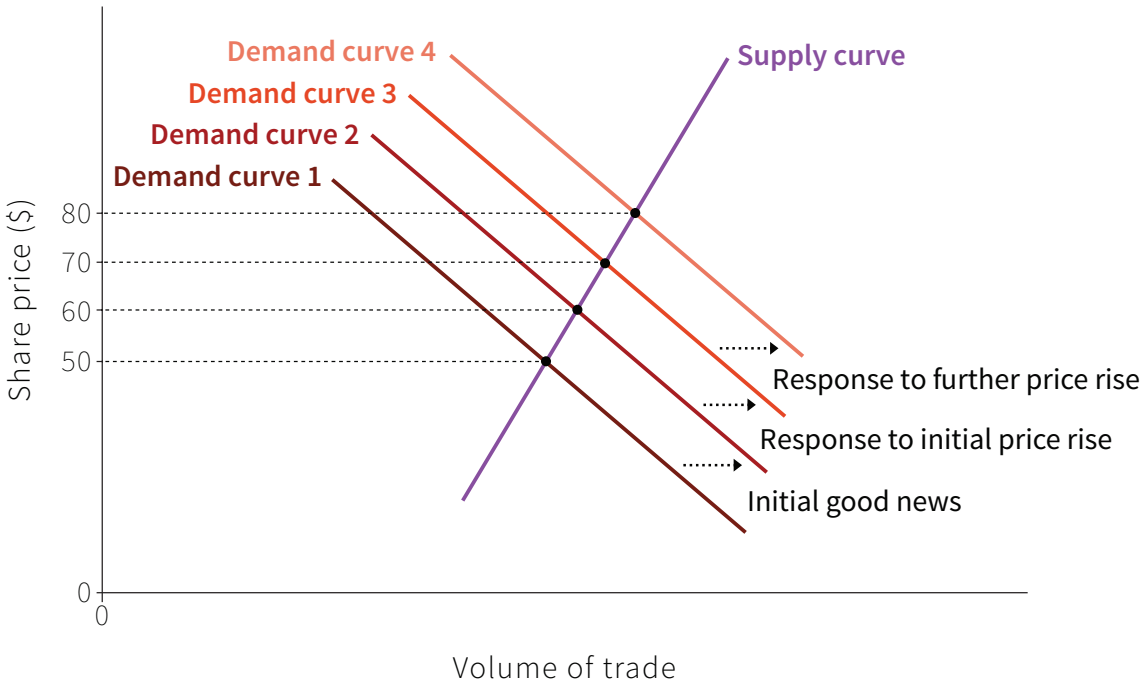


Figure 9.17 *The beginning of a bubble in FCC shares.*

Following the initial price rise, individuals might believe that the price has risen because other people have received news that they hadn't heard themselves, and adjust their own expectations upwards. Or they may think there is an opportunity for speculation: to buy the stock now because they will be able to sell to other buyers at a profit later. Either way, the initial increase in demand created what is called a *positive feedback*, leading to further increases in demand.

This does not happen in the bread market. People do not respond to a rise in bread price by buying more bread and filling their freezer. A different model from Unit 8's is needed for durable assets like shares, paintings or houses. To highlight the role of beliefs in such markets, Figure 9.18 contrasts two scenarios following an exogenous shock of good news about future profits of FCC. In both cases the good news leads to higher demand for shares and the price rises.

In the upper panel, beliefs *dampen* price rises: the response to the higher price is scepticism by some market participants about whether the fundamental value of FCC is really this high ($60 shown in Figure 9.16): as a result, they sell shares, taking a profit from the higher price. This behaviour reduces the price and it falls to a value somewhat above its initial level, where it stabilises. The news has been incorporated into a price between $50 and $60, reflecting the aggregate of beliefs in the market about the new fundamental value of FCC.

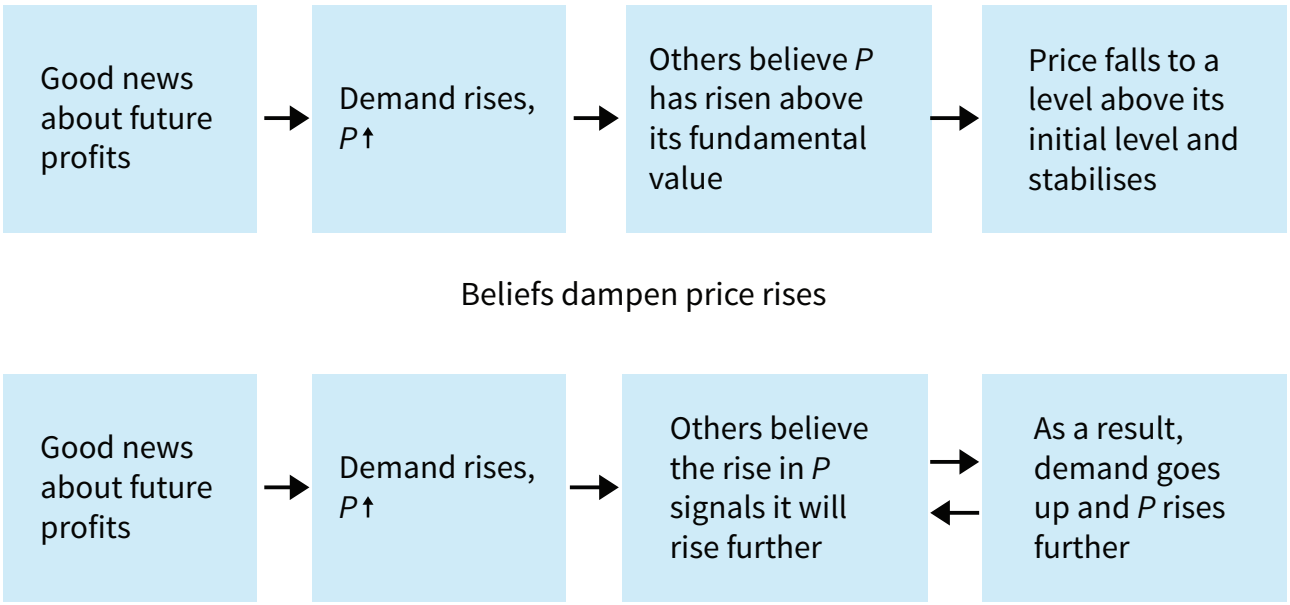


Figure 9.18 *How price changes in asset markets can be dampened through negative feedbacks or amplified through positive feedbacks.*

By contrast, in the lower lower panel beliefs *amplify* price rises. Following the initial price rise, others believe that the rise in price signals that it will rise further. These beliefs produce an increase in the demand for FCC shares. Other traders see that those who bought more shares in FCC benefited as its price rose and so they follow suit. A self-reinforcing cycle of higher prices and rising demand takes hold.

This can be described as a bubble if the price rises significantly beyond the fundamental value of the stock. Note: the self-reinforcing process of rising prices in the lower panel of Figure 9.18 can take place even if there is agreement about the fundamental value of the stock.

The bubble bursts when at least some participants in the market perceive a danger that the price will fall. Then would-be buyers hold back, and those who hold the assets will try to get rid of them. The process in Figure 9.17 is reversed. Figure 9.19 uses the supply and demand model to illustrate what happens. At the top of the bubble the shares trade at $80. Both the supply and demand curves shift when the bubble bursts, and the price collapses from $80 to $54—leaving those who owned shares when the price was $80 with large losses.

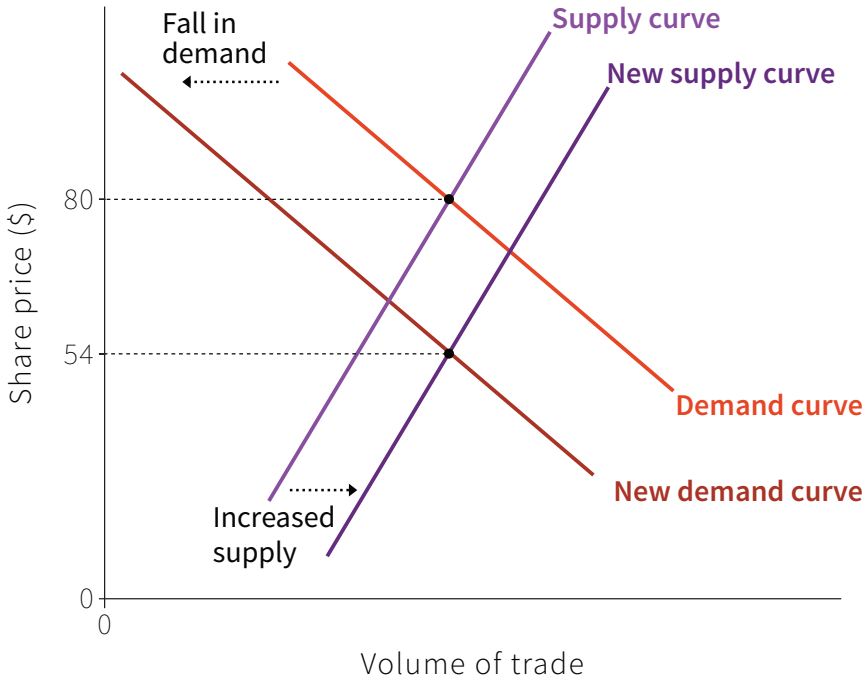


Figure 9.19 *The collapse of the share price in FCC.*

An interesting example of a possible bubble is the recent market for the virtual currency called Bitcoin. Bitcoin was introduced by a group of software developers in 2009. Where it is accepted, it can be transferred from one person to another as payment for goods and services.

It is unlike other currencies in that it is not controlled by a single entity, such as a central bank, but instead is "mined" by individuals who are willing to lend their computing power to verify and record Bitcoin transactions in the public ledger. At the start of 2013 a bitcoin could be purchased for about $13. At its peak on 4 December 2013 the same coin was trading at $1,147. It then lost more than half its value in two weeks. These and subsequent price swings are shown in Figure 9.20.

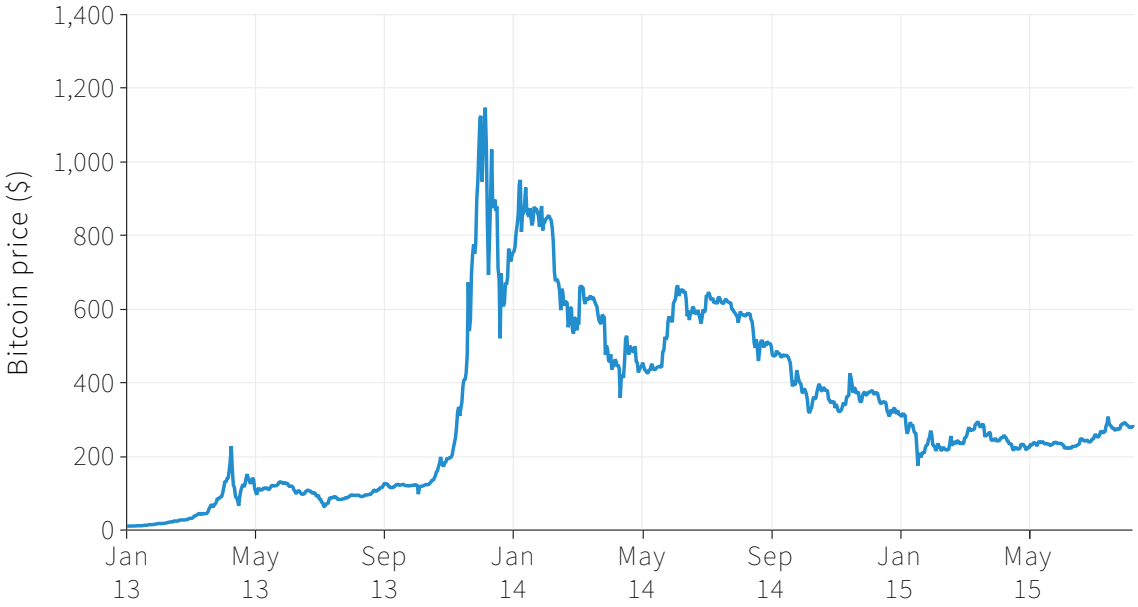


Figure 9.20 *The value of Bitcoin during 2013.*

Source: *Coindesk.com. 2015. 'Bitcoin News, Prices, Charts, Guides & Analysis.' and Bitcoincharts. 2015. Both accessed August.*

If the price of an asset has been driven up solely by beliefs about future price rises, there should be opportunities for those who are well-informed about the value to profit from their superior information. So if the rise in the Nasdaq index was indeed a bubble, why did those who identified it as a bubble fail to profit by placing gigantic bets on a major price decline?

As it happens, many large investors did "lean against the wind" by placing such bets, including some well-known fund managers on Wall Street. The manner in which these bets were placed on the bubble bursting was by *selling short*, or shorting: borrowing shares at the current price believed to be high, and immediately selling them, with the intention of buying them back cheaply (to return to the owner) after the price crashes. The problem is that this is an extremely risky strategy, since it requires accuracy in timing the crash—if prices continue to rise, the losses can become unsustainable. Even if you are right about the bubble, if you get the timing wrong then when you are due to buy the shares and return them to the owner, the price is higher than it was when you sold them. You will make a loss and may not be able to repay your loan.

Indeed, many of those buying an asset may also be convinced of an eventual crash, but are hoping to exit the market before it happens. This was the case during the tech bubble when Stanley Druckenmiller, manager of the Quantum Fund with assets of $8bn, held shares in technology companies that he knew were overvalued. After prices collapsed and inflicted significant losses on the fund, he used a baseball metaphor to describe his error. "We thought it was the eighth inning, and it was the ninth," he explained, "I overplayed my hand."

DISCUSS 9.10: WHAT IS THE FUNDAMENTAL VALUE OF A BITCOIN?

Use the models in this section, and the arguments for and against the existence of bubbles, to provide an account of the data in Figure 9.20.

DISCUSS 9.11: THE BIG TEN ASSET PRICE BUBBLES OF THE LAST 400 YEARS

According to Charles Kindleberger, an economic historian, asset price bubbles have occurred across a wide variety of countries and time periods. The bubbles of the last 100 years have predominantly been focused on real estate, stocks and foreign investment.

1. 1636: The Dutch tulip bubble
2. 1720: The South Sea Company
3. 1720: The Mississippi Scheme
4. 1927–29: The 1920s stock price bubble
5. 1970s: The surge in loans to Mexico and other developing economies
6. 1985–89: The Japanese bubble in real estate and stocks
7. 1985–89: The bubble in real estate and stocks in Finland, Norway and Sweden
8. 1990s: The bubble in real estate and stocks in Thailand, Malaysia, Indonesia and several other Asian countries between 1992 and 1997; and the surge in foreign investment in Mexico 1990–99
9. 1995–2000: The bubble in over-the-counter stocks in the US
10. 2002–07: The bubble in real estate in the US, Britain, Spain, Ireland and Iceland

Pick one of these asset price bubbles, find out more about it, and then:

1. Tell the story of this bubble using the models in this section.
2. Explain the relevance to your story, if any, of the arguments in the *When economists disagree* box about the existence of bubbles.

DISCUSS 9.12: MARKETS FOR GEMS

This article describes how the worldwide markets for opals, sapphires, and emeralds are affected by discoveries of new sources of gems.

1. Explain, using supply and demand analysis, why Australian dealers were unhappy about the discovery of opals in Ethiopia.
2. What determines the willingness to pay for gems? Why do Madagascan sapphires command lower prices than Asian ones?
3. Explain why the reputation of gems from particular sources might matter to a consumer.
4. Shouldn't you judge how much you are willing to pay for a stone by how much you like it yourself?
5. Do you think that the high reputation of gems from particular origins necessarily reflects true differences in quality?
6. Could we see bubbles in markets for gems?

9.9 ECONOMIC RENTS AND THE DYNAMICS OF A CAPITALIST ECONOMY

Throughout this unit, economic rents play an important role in how a capitalist economy works. Recall that an *economic rent* is a payment or other benefit that someone receives that is superior to his or her next best alternative. (In common usage, the word *rent* without the adjective "economic" refers to a payment for the use of housing or a car or land.)

In a private economy rents arise from four sources, each of which you have already studied:

- Innovation (Unit 2)
- Incomplete contracts such as that between an employee and a firm owner (Units 6 and 9)
- Limited competition among buyers or sellers (Unit 7)
- Disequilibrium in a market (this unit)

In this unit we have also seen that economic rents arise when:

- Intervention by government or choices by a firm prevent a market from clearing

Economic rents are of two types: those arising in equilibrium, which are more or less permanent features of the economy, and those that result from some kind of disequilibrium in the economy, which are temporary.

We call equilibrium rents *stationary* because they are persistent. The main examples are shown in Figure 9.21:

TYPE	DESCRIPTION	UNIT
BARGAINING	In a bargaining situation, how much the outcome exceeds the reservation option (next best alternative)	4, 5
EMPLOYMENT	Wages and conditions above an employee's reservation option providing an incentive to work hard	6, 9
MONOPOLY	Profits above economic profits made possible by limited competition	7
CONSUMER SURPLUS	How much more you value the good you purchased than the price you paid. *Next best alternative:* not to have purchased the good.	7, 8
PRODUCER SURPLUS	How much more you receive when you sell the good than the minimum price you would have accepted (marginal cost). *Next best alternative:* not to have produced the good.	7, 8
GOVERNMENT-INDUCED	Payments above the actor's next best alternative not competed away because of government regulation (for example rent control, intellectual property rights)	9

Figure 9.21 *Stationary rents in a capitalist economy.*

By contrast, disequilibrium rents disappear. They set in motion a process—rent-seeking—that ultimately creates an equilibrium in which these kinds of rents no longer exist. For this reason we call them *dynamic*. The main examples are shown in Figure 9.22:

TYPE	DESCRIPTION	UNIT
INNOVATION	The profits above economic profits made possible by being an early innovator. *Next best alternative:* do not innovate	1, 2
DISEQUILIBRIUM PRICE-SETTING	Gains that buyers or sellers receive by changing price or quantity in a market in disequilibrium	9
SPECULATIVE	Profits made by betting correctly on price changes of durable assets during bubbles	9

Figure 9.22 *Dynamic rents in a capitalist economy.*

Economic rents and rent-seekers often have a bad name in economics. People disapprove because they think about rents as those arising from government-created monopolies (taxi licenses, intellectual property rights) or privately-created monopolies. These rents indicate that the good or service will be sold at a price exceeding its marginal cost, and so the markets for these goods are not Pareto efficient.

But in this and earlier units we have seen the usefulness of some economic rents:

- *Innovation rents*: The prospect of earning innovation rents was central to the permanent technological revolution that was part of the capitalist revolution.
- *Employment rents*: The cost of job loss is a device that the employer uses to ensure that the worker will work up to the expected standard; in its absence little would be produced at all.
- *Disequilibrium rents*: At the beginning of this unit we saw how mobile phones allowed the Kerala fishermen to be effective rent-seekers, which substantially enhanced the efficiency of the sardine market: rent-seeking brought an out-of-equilibrium market into equilibrium, when excess supply or demand created opportunities for buyers and sellers to benefit by changing their prices or the quantities transacted.

Economic rents are closely associated with situations in which one or more actors has the ability to benefit from setting the price or wage (or some other decision) rather than just replicating what others are doing, for example by assuming the current wage, or the price of a good, to be fixed. We have seen Units 8 and 9 that when a competitive market is in equilibrium there are no rents to be had by setting a price different from what others are doing. But Figure 9.23 makes it clear that acting as a price- or wage-setter is often a profitable way to behave when there are rents to be captured or created:

SITUATION	ACTION	REASON	UNIT
Equilibrium of a competitive market	Price-taking	No rent-seeking possible	8
Disequilibrium of a competitive market	Price-setting	To exploit dynamic rents	9
Firms with limited competitors	Price-setting	To ensure permanent monopoly rents	7
Firms hiring employees	Wage-setting	To provide permanent rents to motivate employees	6

Figure 9.23 *Price-taking and price-setting.*

9.10 CONCLUSION

The capitalist economy combines both the decentralised decision-making of the market, illustrated by the chain of events triggered by the American civil war and rise in the price of cotton, with the centralised decision-making process in large firms. The decision by the owners and managers of Dobson and Barlow to develop new machinery for textile mills suitable for Indian cotton was not implemented through price messages, but by orders to the company's engineers and mechanics to undertake the work.

The balance of these two systems—decentralised and centralised—in a capitalist economy shifts over time, as we saw in Unit 6, as firms decide to outsource production or to take on the production of parts previously acquired by purchase. Where massive changes in the use of a society's resources need to happen quickly, such as in wartime, virtually all economies resort to planning—

CONCEPTS INTRODUCED IN UNIT 9

Before you move on, review these definitions:

- *Long-run* and *short-run equilibria*
- *Market equilibration through rent-seeking*
- *Profit curve*
- *Continuous double price auction*
- *Order book*
- *Price bubble*
- *Fundamental value of a share*
- *Positive feedback*
- *Dynamic and stationary economic rents*

as the UK and the US did in the second world war. But for the normal changes in an economy the use of prices as messages works well, as illustrated by the case of cotton prices.

Sometimes suppliers or regulators choose to override price messages, leading to persistent excess supply or demand, as we have seen in the cases of concert tickets and housing rents. But many market prices are free to change when conditions of supply or demand change, and markets represent a flexible way to inform members of an economy of the relative scarcity of goods, giving them a reason (their own desire to save or make money) to respond in a way that makes better use of an economy's productive capacity.

But not all prices send the right message. We have already seen how the wrong messages are sent when bubbles develop. In the next unit we describe the conditions under which markets send the right messages, and the reasons why they sometimes fail to do so.

Key points in Unit 9

Rent-seeking activities

Rent-seeking is made possible by market disequilibrium.

Disequilibrium rents may help to clear a market

In some (but not all) markets this process eliminates excess supply or demand.

Bubbles and crashes

In some markets—for example for financial assets—rent-seeking may lead the price to deviate from its fundamental value, creating a bubble or a crash. This could happen if traders believe that future price changes will be in the same direction as current price changes, due to positive feedbacks.

The labour market does not clear

In the labour market the long-run equilibrium wage, and employment, is determined by the wage curve and the profit curve. In equilibrium there is unemployment, because of the way profit-maximising firms set wages.

Continuous double auctions

In markets for financial assets, supply and demand shift as traders receive new information. The price adjusts in a continuous double auction to reconcile supply and demand.

Financial markets

Financial markets are subject to rapid fluctuations in prices because:

- Financial assets are often purchased for their resale value.
- Large transactions are possible due to the ease of borrowing.
- Prices depend greatly on how people's expectations of how other people will value the asset.

Price-setting may create rents

Firms or governments may choose to set a price that does not clear the market, giving rise to excess demand or supply, and potential economic rents.

Rents are often useful

Dynamic and stationary rents are essential to the functioning of a modern capitalist economy.

9.11 EINSTEIN

The profit curve

There are several steps to show how the profit curve for the economy as a whole results from the decisions of individual firms.

Step 1: The firm sets its price

We simplify by assuming the firm's only costs are the wages it pays, and that on average a worker produces λ units of output, which does not vary with the number of workers employed. (We use the Greek letter lambda, λ, for labour productivity.) The firm pays the worker a wage W in dollars. Both labour productivity and wages can be measured per hour, per day or per year. In our numerical examples, we typically use hourly wages and productivity.

The unit labour cost is the wages to hire the labour sufficient to produce one unit of the good. This is:

$$unit\ labour\ cost = \frac{nominal\ wage}{labour\ productivity} = \frac{W}{\lambda}$$

For example: if W = \$30 and λ = 10, then unit labour cost is \$30/10 units, which equals \$3 per unit.

Recall from Unit 7 that the firm chooses its price so that the markup is inversely proportional to the elasticity of the demand curve it faces:

$$\frac{(price - marginal\ cost)}{price} = \frac{1}{elasticity}$$

The elasticity of the firm's demand curve depends on how much competition the firm faces from other firms. So the higher the elasticity, the lower the firm's price and markup. In other words, the inverse of the demand elasticity is a measure of the competition that the firm faces, which we will call μ (the Greek letter mu, rhyming with "few"):

$$\mu = \frac{1}{elasticity}$$

Using our assumptions, the firm's marginal (and average) cost is its unit labour cost, W/λ, and we can say that the firm sets its price p so that:

$$\mu = \frac{p - \left(\frac{W}{\lambda}\right)}{p}$$

Which gives:

$$\mu = 1 - \frac{(W/p)}{\lambda}$$

Rearranging, we get:

$$\frac{(W/p)}{\lambda} = 1 - \mu$$

Multiply each side by λ:

$$\frac{W}{p} = \lambda(1 - \mu) = \lambda - \lambda\mu$$

We call this *Equation 1*. In words, it says:

$$\frac{W}{p} = output\ per\ worker - real\ profit\ per\ worker$$

When the firm sets its profit-maximising price, this splits output per worker in the firm into a part that goes to employees as wages, and a part that goes to owners as profits.

Step 2: The price level in the economy as a whole, and the real wage

From the employee's point of view, the real wage measures how much of her typical consumption she can purchase with an hour's earnings. Since she buys many different goods and services, this depends on prices set by the firms throughout the economy, not just her own firm. Suppose the entire economy is made up of firms facing competition conditions similar to the single firm we have just studied. We call the average price of the goods and services the worker consumes, P, which is an average of the different levels of p set by individual firms across the economy.

The real wage is the nominal wage divided by the economy-wide price level, P:

$$\begin{aligned} \textit{real profit} &= \frac{\textit{nominal wage}}{\textit{price level}} \\ &= \frac{W}{P} \\ &= w \end{aligned}$$

Step 3: Profits, wages, and the profit curve

We can use *Equation 1* from Step 1 to see the economy-wide real wage that will result in the markup consistent with the extent of competition facing the firms. Across the entire economy:

$$\frac{W}{P} = \lambda(1-\mu) = \lambda - \lambda\mu$$

In words:

$$\textit{real wage} = \textit{output per worker} - \textit{real profit per worker}$$

This is the wage indicated by the profit curve.

9.12 READ MORE

Bibliography

1. Arnott, Richard. 1995. 'Time for Revisionism on Rent Control?' *Journal of Economic Perspectives* 9 (1): 99–120.
2. Bosvieux, Jean, and Oliver Waine. 2012. 'Rent Control: A Miracle Solution to the Housing Crisis?' *Metropolitics*. November 21.
3. Brunnermeier, Markus. 2009. 'Lucas Roundtable: Mind the Frictions.' *The Economist*. August 6.
4. Campbell, Gareth. 2012. 'Myopic Rationality in a Mania.' *Explorations in Economic History* 49 (1): 75–91.
5. Cassidy, John. 2010. 'Interview with Eugene Fama.' *The New Yorker*. January 13.
6. Garber, Peter M. 1989. 'Tulipmania.' *Journal of Political Economy* 97 (3): 535.
7. Giberson, Michael. 2010. 'I Cringe When I See Hayek's Knowledge Problem Wielded as a Rhetorical Club.' *Knowledge Problem*. April 5.
8. Gomelsky, Victoria. 2014. 'On the Origins of Gems.' *New York Times Fashion and Style*, March 16.

9. Harford, Tim. 2012. 'Still Think You Can Beat the Market?' *The Undercover Economist*. November 24.
10. Hayek, Friedrich A. 1945. 'The Use of Knowledge in Society.' *The American Economic Review* XXXV (4): 519–30.
11. Hayek, Friedrich A. 1948. *Individualism and Economic Order*. Chicago, Il: University of Chicago Press.
12. Hayek, Friedrich A. (1945) 1994. *The Road to Serfdom*. Chicago, Il: University of Chicago Press.
13. Hayek, Friedrich A. (1946) 2010. 'The Meaning of Competition (Stafford Little Lecture, Princeton University).' *Mises Institute*.
14. Jensen, Robert. 2007. 'The Digital Provide: Information (Technology), Market Performance, and Welfare in the South Indian Fisheries Sector.' *The Quarterly Journal of Economics* 122 (3): 879–924.
15. Kestenbaum, David. 2011. 'Ranking Cute Animals: A Stock Market Experiment.' *NPR.org*. January 14.
16. Keynes, John Maynard. 1936. *The General Theory of Employment, Interest and Money*. London: Palgrave Macmillan.
17. Kindleberger, Charles P. (1978) 2005. *Manias, Panics, and Crashes: A History of Financial Crises* (Wiley Investment Classics). Hoboken, NJ: Wiley, John & Sons.
18. List, John A. 2004. 'Testing Neoclassical Competitive Theory in Multilateral Decentralized Markets.' *Journal of Political Economy* 112 (5): 1131–56.
19. Lucas, Robert. 2009. 'In Defence of the Dismal Science.' *The Economist*. August 6.
20. Malkiel, Burton G. 2003. 'The Efficient Market Hypothesis and Its Critics.' *Journal of Economic Perspectives* 17 (1): 59–82.
21. Shiller, Robert J. 2003. 'From Efficient Markets Theory to Behavioral Finance.' *Journal of Economic Perspectives* 17 (1): 83–104.
22. Shiller, Robert J. 2015. *Irrational Exuberance*. Chapter 1. Princeton, NJ: Princeton University Press.
23. Smith, Vernon L. 1962. 'An Experimental Study of Competitive Market Behavior.' *Journal of Political Economy* 70 (3): 322.
24. Smith, Vernon L. 1994. 'Economics in the Laboratory.' Journal of Economic Perspectives 8 (1). *American Economic Association*: 113–31.
25. Stiglitz, Joseph E. 1990. 'Symposium on Bubbles.' *Journal of Economic Perspectives* 4 (2): 13–18.
26. *The Economist*. 2013. 'Was Tulipmania Irrational?' October 4.
27. *The Economist*. 2014. 'Keynes and Hayek: Prophets for Today.' March 14.

10

MARKETS, CONTRACTS AND INFORMATION

Courtesy of US Coastguard

HOW ADAM SMITH'S "INVISIBLE HAND" MAY FAIL, AND HOW PRIVATE BARGAINING AND GOVERNMENT POLICY SOMETIMES IMPROVE THE OUTCOME

- Governments provide essential conditions in which markets can exist and work well, including private property rights and enforcement of contracts
- Market failures are Pareto-inefficient allocations in which potential mutual benefits from an exchange or other economic interaction are not realised, which will occur when markets are not competitive
- Even when markets are competitive, market failures may occur if economic actors do not take full account of the effect of their actions on others
- This will be the case when some aspect of an exchange (including effects on those not involved in it) is not covered by property rights and contracts that can be enforced
- Costly environmental spillovers and the positive effects of knowledge creation by R&D are examples of external effects that are not fully covered in contracts
- The reason is that the information necessary to enforce the necessary rights and contracts is not available to one or more of the parties
- Private bargaining, government policy, or a combination of the two may improve a market allocation when there are external effects
- For moral and political reasons some goods and services (for example our vital organs or our votes) are not traded on markets, but are allocated by other means

Institute for New Economic Thinking
UCL SciencesPo Azim Premji University
FRIENDS PROVIDENT Foundation

Pick up any object in reach, and ask this thing a few questions:

You Thing, who made you, and do I know your maker?
Do you come from a place that I've ever been to?
Where do the materials or parts that you're made of come from?

Maybe you picked up your phone, your cup or your mouse. Whatever you picked up, the thing's answers (if things could talk) would be the same:

Thing I was made by people you have not met, and will probably never meet, living in places you will probably never visit.

You So how did you end up on my desk?

Thing Long story. You are busy studying, so I'll just give you a very short answer.

You How short?

Thing *The market*.

Now imagine that the same questions had been asked of something anywhere in the world the year that Adam Smith wrote *The Wealth of Nations*. Many of the things that you might have picked up in 1776 would have been made by a member of the family, or of the village. Some would have been within reach because you had made the object yourself, some because you had use of them as a family member, and others would have been purchased from neighbours.

One of the changes that was underway during Adam Smith's life, but has greatly accelerated since, is specialisation in the production of the goods and services around us. As Smith explained, we become better at producing things when we each focus on a limited range of activities. The same is true of firms that often produce at lower unit cost by producing a large number of identical goods, as we have seen in Unit 7.

But people will not specialise unless they have a way to acquire the other goods on which their livelihood depends. That is where the market comes in. Chapter 3 in Smith's *Wealth of Nations* is called: "That the Division of Labour is Limited by the Extent of the Market":

"When the market is very small, no person can have any encouragement to dedicate himself entirely to one employment, for want of the power to exchange all that surplus part of the produce of his own labour, which is over and above his own consumption, for such parts of the produce of other men's labour as he has occasion for."
— Adam Smith, *An Inquiry into the Nature and Causes of the Wealth of Nations* (1776)

The market, alongside firms, as we saw in Unit 6, is what makes specialisation and the division of labour possible. And as the extent of markets expanded both globally and also to encompass goods and services that in the past were distributed within families, or as gifts, or not exchanged at all, specialisation and the division of labour has expanded with it. Today, the American retailer Walmart sells more than 4 million different products.

When you think about markets, the word that often comes to mind is "competition". But markets are also the foundation of the largest cooperative venture that our species has ever undertaken, in which billions of us engage, for the most part unwittingly, in providing the goods and services on which others live, gaining our livelihoods in return.

The market does more than allow the extension of the division of labour. Markets are a way of governing an economy.

In *The Wealth of Nations*, Adam Smith explained how the owners of capital (motivated by their individual desire for profit) and others (through their pursuit of a more comfortable or pleasant life) would make economic decisions that would benefit society as a whole. Capital would be invested where it was most productive, and the consumption of goods and services would economise on society's scarce resources. He wrote that each individual could be:

> "[L]ed by an invisible hand to promote an end [the well-being of others] which was no part of his intention."
> — Adam Smith, *An Inquiry into the Nature and Causes of the Wealth of Nations* (1776)

In Unit 9, we examined Friedrich Hayek's explanation of how Smith's invisible hand could work. When it does, Hayek explained, prices send messages about the real scarcity of goods and services, messages that motivate people to produce, consume, invest, and innovate in ways that result in the best use being made of an economy's productive potential. Hayek suggested we think of the market as a giant information-processing machine, providing information that guides the economy, usually in desirable directions.

The remarkable thing about this massive computational device is that it's not really a machine at all: nobody designed it; nobody is at the controls. When it works well we use phrases like "the magic of the market."

But sometimes the magic fails. In this unit we will ask how well the market works, and will consider cases in which prices send the wrong messages. Smith explained that in many areas, such as education and the legal system, government policies were needed to promote social wellbeing and to ensure that markets work well. Smith was also clear that that there were some things that should not be bought and sold on markets. The modern equivalents might include human kidneys, votes, a good school or life-saving medical care, and you might believe that some of these things should be allocated in some other way.

For both Smith and Hayek, competition among many buyers and sellers is an essential part of the magic of the market and, when it is absent or limited, the invisible hand will not work.

We examined the implications of competition in Units 7 and 8:

- Firms facing little competition—monopolists or those producing differentiated goods—set their prices above marginal cost.
- *The price at which the good is sold then sends the wrong message*: The high price overstates the real scarcity of the good as indicated by its marginal cost.
- *The resulting allocation is not Pareto efficient*: Too little is sold, so there is a deadweight loss.
- *In contrast, firms in competitive markets are price-takers*: They produce where price is equal to marginal cost, and the allocation maximises the total surplus of the buyers and sellers.

For a market to work well (or even to exist) other social institutions and social norms are required. Governments, for example, provide a system of laws and law enforcement that allow markets to function effectively, by guaranteeing *property rights* and enforcing *contracts*. *Social norms* dictate that you respect the property rights of others, even when enforcement is unlikely or impossible.

As we saw in Unit 1, if something is to be bought and sold then it must be possible to claim the right to own it. A purchase is simply a transfer of ownership rights from the seller to the buyer. You would hesitate to pay for something unless you believed that others would acknowledge (and if necessary protect) your right to keep it.

And whenever you agree with a seller to pay a certain amount of money in exchange for a good—say a pair of shoes—you implicitly enter into a contract with the seller. If you have the protection of a legal system you can expect the contract to be honoured: when you get home and open the box the shoes will be there, and if they fall apart within days you will receive a refund. It is the government that determines the trading rules—the rules of the game in which market trade takes place. Of course enforcement by a court is rarely necessary because of social norms that motivate both buyers and sellers to play by the rules of the game, even in cases where there is not literally a contract or a transfer of a title of ownership.

More complex transactions require explicit written contracts that can be used in court as evidence that the parties agreed a transfer of ownership. For example, an author may sign a contract giving a publisher the sole right to publish a book. Contracts govern relationships that are to be maintained over a period of time, particularly employment: in the labour market, a court upholds the right of the worker to work no more than contracted hours and to receive the agreed upon pay.

Laws and legal traditions can also help markets function when they provide compensation for individuals who are harmed by the actions of others. Liability law, for example, ensures that if a firm sells a car with a design fault, and someone is injured as a result, the firm must pay for the damage. Employers usually have a duty of care towards their employees, requiring them to provide a safe working environment, and incurring fines or other penalties when they do not.

But providing conditions that facilitate trade may not be enough for a market to work well.

Since the discovery of penicillin in 1928, the development of antibiotics has brought huge benefits to mankind. Diseases that were once fatal are now easily treatable with medicines that are cheap to produce. But the World Health Organisation has recently warned that we are heading for a "post-antibiotic era" as bacteria are becoming resistant: "Unless we take significant actions to... change how we produce, prescribe and use antibiotics, the world will lose more and more of these global public health goods and the implications will be devastating."

Overuse of antibiotics is an example of a *social dilemma* studied in Unit 4 when the unregulated pursuit of self interest leads to outcomes that are Pareto inefficient. Bacteria become resistant to antibiotics when we use them too often, in the wrong dosage or for conditions that are not caused by bacteria. In India, for example, antibiotics are easily available over the counter in pharmacies without a doctor's prescription.

Doctors recognise that leaving the allocation of antibiotics to the market is having damaging consequences. On the advice of unlicensed private medical practitioners, people use antibiotics when other treatments would be better. To save money, the patients often stop taking the antibiotics when they feel a little better. This is exactly the pattern of use that will produce antibiotic-resistant pathogens. But, for the patient, the treatment worked, and the unlicensed doctor's business will prosper.

When the market allocation of a good or service is Pareto inefficient, we have what is called a *market failure*. In the case of antibiotics this happens because the decisions of the buyers and sellers also have costs or benefits for other people that the decision makers do not take into account. Future users of antibiotics are put at risk from resistant bacteria. Such an effect is known as an external effect. In the next section we look at another external effect: the case of pollution.

MARKET FAILURE

This occurs when markets allocate resources in a Pareto-inefficient way.

DISCUSS 10.1: PROPERTY RIGHTS AND CONTRACTS IN MADAGASCAR

Marcel Fafchamps and Bart Minten studied grain markets in Madagascar in 1997, where the legal institutions for enforcing property rights and contracts were weak. Despite this, they found that theft and breach of contract were rare. The grain traders avoided theft by keeping their stocks very low and, if necessary, sleeping in the grain stores. They refrained from employing additional workers for fear of employee-related theft. When transporting their goods they paid protection money and travelled in convoy. Most transactions took a simple "cash and carry" form. Trust was established through repeated interaction with the same traders.

1. Do these findings suggest that strong legal institutions are not necessary for markets to work?
2. Consider some market transactions in which you have been involved. Could these markets work in the absence of a legal framework, and how would they be different if they did?
3. Can you think of any examples of transactions where repeated interaction helps to facilitate market transactions?
4. Why might this be important even when a legal framework is present?

10.1 MARKET FAILURE: EXTERNAL EFFECTS OF POLLUTION

When we analyse the gains from trade in markets for consumer goods such as cars, books, clothes or washing machines using the methods in Units 7 to 9, we measure the gains to the buyers and sellers using consumer and producer surplus. But it is not enough that all of the potential consumers' or producers' surpluses should be realised when we want a market to work well. Why not? We have to consider also the costs or benefits that may be experienced by people who are neither buyers nor sellers but are somehow affected. The superbug that emerges as a result of the sale and overuse of an antibiotic, for example, may kill someone who had no part in the purchase and sale of the antibiotic.

So any evaluation of how well markets work must include the costs or benefits to others affected by the consumption or production of the good. We will use this approach to analyse the case in which the production of a good creates an external cost: pollution.

In the Caribbean islands of Guadeloupe and Martinique (both part of France), the pesticide Chlordecone was used on banana plantations from 1972 until 1993 to kill banana weevil, reducing costs and boosting the plantations' profits. As the chemical was washed off the land into rivers that flowed to the coast it contaminated freshwater prawn farms, the mangrove swamps where crabs were caught, and coastal fisheries.

To see why this is called an external effect, imagine for a minute that the banana plantations and fisheries were owned by the same large company that hired fishermen and sold what they caught for profit. The owners of the company would decide on the level of pesticide to use, taking account of its downstream effects. They would trade off the profits from the banana part of their business against the losses from the fisheries.

But this was not the case. The profits from banana production, which were increased by using pesticide, were owned by the plantations. The losses from fishing were owned by the fishermen. The pollution effect of the pesticide was external to the people making the decision on its use. Joint ownership of the plantations and fisheries would have internalised this effect. In Martinique and Guadeloupe the plantations and fisheries were under separate ownership.

To investigate the implications of this kind of external effect, Figure 10.1 shows the marginal costs of growing bananas on an imaginary Caribbean island where a fictional pesticide called Weevokil is used. The marginal cost of producing bananas for the growers is labelled as the *marginal private cost* (MPC). It slopes upward because the cost of an additional tonne increases as the land is more intensively used, requiring more Weevokil. Use the slideline to compare the MPC with the *marginal social cost* (MSC), which takes into account the costs borne by fishermen whose waters are contaminated by Weevokil.

You can see in Figure 10.1 that the marginal social cost of banana production is higher than the marginal private cost. To focus on the essentials, we will consider a case in which the wholesale market for bananas is competitive, and the market price is $400 per tonne. Then, if the banana plantation owners wish to maximise their profit, we know that they will choose their output so that price is equal to marginal cost—that is, marginal private cost. Figure 10.2 shows that their total output will be 80,000 tonnes of bananas, at point A. Although 80,000 tonnes maximises profits for banana producers, this does not include the cost imposed on the fishing industry, so it is not a Pareto-efficient outcome.

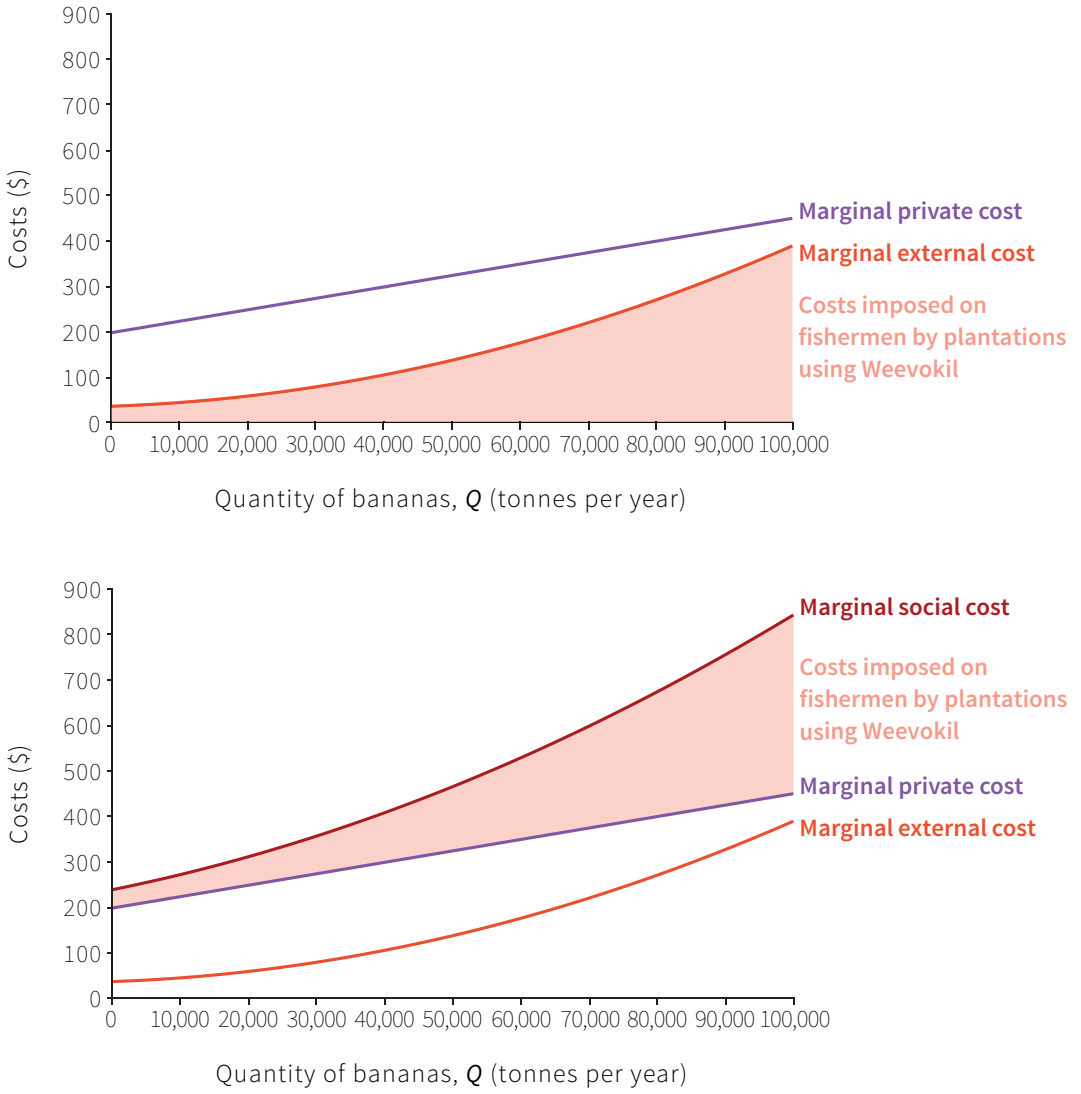


Figure 10.1. *Marginal costs of banana production using Weevokil.*

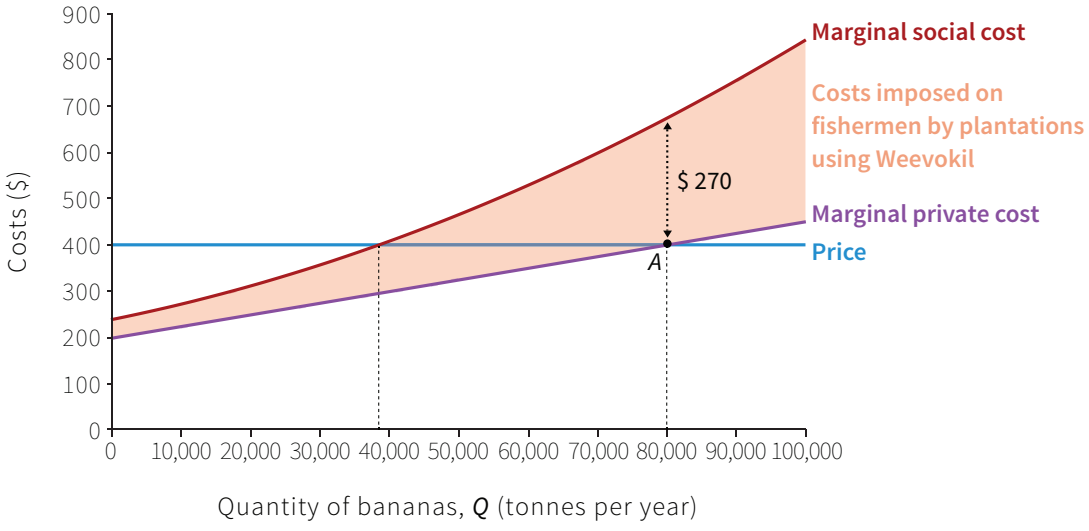


Figure 10.2. *The choice of banana output.*

To see this, think about what would happen if the plantations were to produce less. The fishermen would benefit but the owners of the plantations would lose. So on the face of it, it appears that producing 80,000 tonnes must be Pareto efficient. But let's imagine hypothetically for now that the fishermen could pay the plantation owners to produce one tonne less. The fishermen would gain $270—they would no longer suffer the loss of revenue from fishing that is caused by the production of the 80,000th tonne of bananas. The plantations would lose hardly anything (their revenues would fall by $400) but their costs would fall by almost exactly this amount because, when producing 80,000 tonnes, the marginal private cost is equal to the price ($400). Thus if a payment from the fishermen to the plantation owners could be arranged for any amount between just greater than zero and just less than $270, both groups would be better off with 79,999 tonnes of bananas.

What about another payment to get the plantations to produce 79,998 tonnes instead? You can see that, because the marginal external cost imposed on the fishermen is still much higher than the price received by the plantations, such a payment would also make both parties better off.

Where would this hypothetical process come to an end?

Look at the point in Figure 10.2 at which the price of bananas is equal to the marginal social cost. At this point, 38,000 tonnes of bananas are produced. If the payments by the fishermen to the plantations resulted in them producing just 38,000 tonnes, then the fishermen could no longer benefit by making further payments in return for reduced output. If production were lowered further, the loss to the plantations (the difference between price and marginal cost) would be greater than the gain to the fishermen (the difference between private and social cost, shaded). At this point, the maximum payment the fishermen would be willing to make would not be enough to induce the plantations to cut production further. So 38,000 tonnes is the Pareto-efficient level of banana output.

To summarise:

- The plantations produce 80,000 tons of bananas, at which price equals MPC.
- The Pareto-efficient level of output is 38,000 tonnes of bananas, where price equals MSC.
- When production is 38,000 tonnes it is not possible for the plantations and fishermen, jointly, to be made better off.
- If the banana plantations and fisheries were owned by a single company, it would choose to produce 38,000 tonnes; for the single owner, price would be equal to MPC at 38,000 tonnes.

We can summarise the problem of a negative external effect by looking at the decision taken, its effect on others, whether this is a cost or benefit and the market failure that results. We have laid this out in the the table below. The terms that are used to label the market failure are in the final column.

Throughout this unit we use this method to consider different cases of market failure, and summarise each example in a similar table. In the conclusion we will bring all the small tables together in Figure 10.11 so that you can compare them more easily.

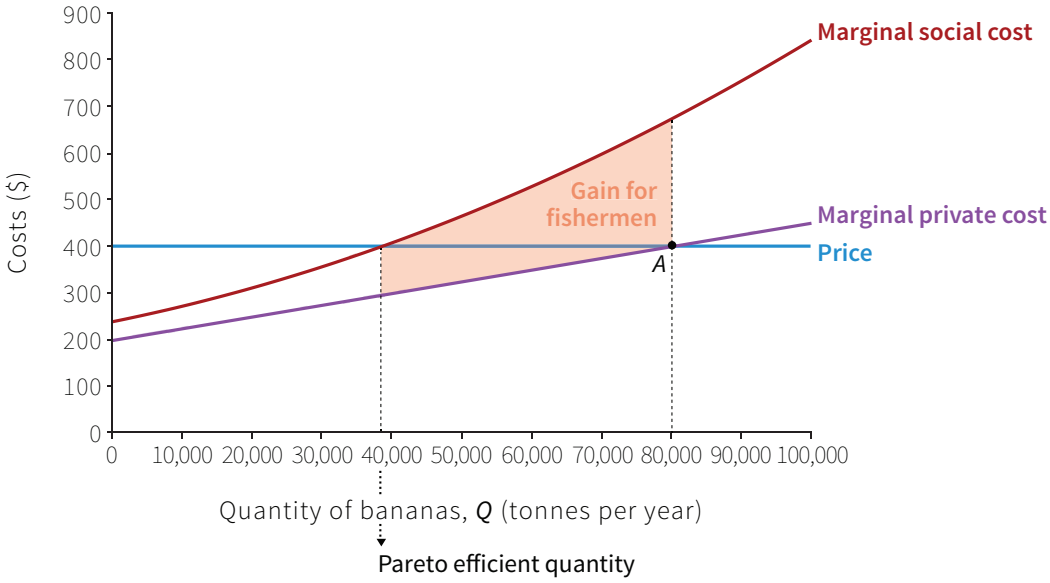
THE DECISION	HOW IT AFFECTS OTHERS	COST OR BENEFIT	MARKET FAILURE (MISALLOCATION OF RESOURCES)	TERMS APPLIED TO THIS TYPE OF MARKET FAILURE
A firm uses a pesticide that runs off into waterways	Downstream damage	Private benefit, external cost	Overuse of pesticide and overproduction of crop in which it is used	Negative external effect, environmental spillovers

10.2 EXTERNAL EFFECTS AND BARGAINING

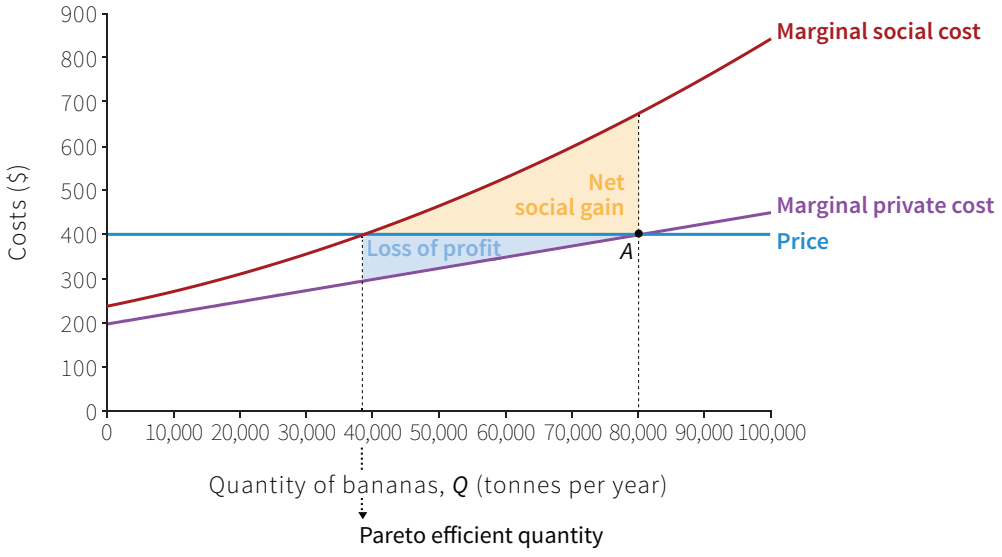
Hypothetically the fishermen could pay the plantation owners to produce fewer bananas, and this shows that the market allocation (producing 80,000 tonnes of bananas) is not Pareto efficient. But does it suggest a remedy for this market failure that could be implemented in the real world?

Let's see how a private bargain might solve the problem. Initially it is not illegal to use Weevokil: the plantations have the right to use it, and they produce 80,000 tonnes of bananas. This allocation and the incomes, environmental effects, and other outcomes represent the reservation position of the plantation owners and fishermen. This is what they will get if they do not come to some agreement.

For the fishermen and the plantation owners to negotiate, they would each have to be organised so that a single person (or body) could make agreements on behalf of the entire group. So let's imagine that a representative of an association of fishermen sits down to bargain with a representative of an association of banana growers. To keep things simple we will assume that, at present, there are no feasible alternatives to Weevokil; so they are bargaining over the output of bananas.



The situation before bargaining begins is represented by point A, and the Pareto efficient quantity of bananas is 38,000 tonnes. The gain for fishermen (from cleaner water) if output is reduced from 80,000 to 38,000 is shown by the shaded area.



The net social gain is the gain for the fisherman minus the loss for the plantations, shown by the remaining yellow area.

Figure 10.3. *The gains from bargaining.*

Both sides should recognise that they could gain from an agreement to reduce output to the Pareto-efficient level. In Figure 10.3 the situation before bargaining begins is point A, and the Pareto-efficient quantity is 38,000 tonnes. The gain for the fishermen (from cleaner water) if output is reduced from 80,000 to 38,000 is shown by the total shaded area. But reducing banana production will lead to lower profits for the plantations. Use the slideline to see that the fall in profit is smaller than the gain for the fishermen, so there is a net social gain that they could agree to share.

Since the gain to the fisherman would be greater than the loss to the plantations, the fishermen would be willing to pay the banana growers to reduce output to 38,000 tonnes *if they had the funds to do so*.

The *minimum acceptable payment* is determined by what the plantations get in the existing situation: their reservation profits (shown by the blue area labelled "loss of profit"). If this minimum payment to compensate the plantation owners for their loss of profit were the deal they struck, the fishing industry would achieve a net gain from the agreement equal to the net social gain, while plantations would be no better (and no worse) off.

The *maximum* the fishing industry would pay is determined by their *fallback (reservation) option*, as in the case of the plantations. It is the sum of the blue and yellow areas; in this case the plantations would get all of the net social gain, while the fishermen would be no better off. Unit 5 showed us that the compensation they agree on, between these maximum and minimum levels, will be determined by the bargaining power of the two groups.

You may think it unfair that the fishermen need to pay for a reduction in pollution. At the Pareto-efficient level of banana production the fishing industry is still suffering from pollution, and it has to pay to stop the pollution getting worse. This happens because we have assumed that the plantations have a legal right to use Weevokil.

An alternative legal framework could give the fishermen a right to clean water. If that were the case, the plantation owners wishing to use Weevokil could propose a bargain in which they paid the fishermen to give up some of their right to clean water to allow the Pareto-efficient level of banana production, which will be a much more favourable outcome for the fishermen. In principle the bargaining process would result in a Pareto-efficient allocation independently of whether the initial rights were granted to the plantations (right to pollute) or to the fishermen (right to unpolluted water). But the two cases differ dramatically in the distribution of the benefits of solving the market failure.

And, more important for the question of Pareto efficiency of market allocations, in practice there are always obstacles to bargaining:

- *Impediments to collective action*: Private bargaining may be impossible if there are many parties on both sides of the external effect, many fishermen and many plantation owners in our example. Each side needs to find someone they trust to bargain for them, and agree how payments will be shared within each industry. The individuals representing the two groups would be performing a public service that might be difficult to secure.
- *Missing information*: Devising the payment scheme makes it necessary to measure the costs of Weevokil, not just in aggregate, but to each fisherman. We also need to establish the exact origin of the pollutant, plantation by plantation. Only when we have this information can we calculate the size of the payment that each fisherman has to pay, and how much that each plantation should receive. It's easy to see that it is far harder to make a polluting industry accountable for the damage it does than to calculate the liability for damage done, for example, by a single reckless driver.
- *Enforcement*: The bargain must be enforceable. Having agreed to pay thousands of dollars, the fishermen must be able to rely on the legal system if a plantation owner does not reduce output as agreed. This may require the fishermen and the courts to discover information about the plantation's operations that are not public.
- *Limited funds*: The fishermen may not have enough money (we will see why they would probably not be able to borrow large sums in the next unit) to pay the plantations to reduce output to 38,000 tonnes.

The bargaining approach to resolving market failures was developed by Ronald Coase, and to honour his contribution economists refer to the fishermen's attempt at a private settlement with the plantation owners as Coasian bargaining.

The pesticide example illustrates that, although ideally correcting market failures through bargaining may not require direct government intervention, it does require a legal framework for enforcing contracts so that property rights are tradable and so that all parties stick to the bargains they make. Even with this framework the problems of collective action, missing information and enforcement of what will inevitably be complex contracts make it unlikely that Coasian bargaining alone can address market failures.

DISCUSS 10.2: BARGAINING POWER

In the example of plantation owners and fishermen, can you think of any factors that might affect the bargaining power of these parties?

GREAT ECONOMISTS

RONALD COASE

Ronald Coase (1910-2013) had the insight to argue that when one party is engaged in an activity that has the incidental effect of causing damage to another, a negotiated settlement between the two may result in a Pareto-efficient allocation of resources. He used the legal case of Sturges v Bridgman to illustrate his argument. The case concerned Bridgman, a confectioner (candy-maker) who for many years had been using machinery that generated noise and vibration. This caused no external effects until his neighbour Sturges built a consulting room on the boundary of his property, close to the confectioner's kitchen. The courts granted the doctor an injunction that prevented Bridgman from using his machinery.

Coase pointed out that, once the doctor's right to prevent the use of the machinery had been established, the two sides could modify the outcome. The doctor would be willing to waive his right to stop the noise in return for a compensation payment. The confectioner would be willing to pay if the value of his annoying activities exceeded the costs that they imposed on the doctor. Also, the court's decision would make no difference to whether Bridgman continued to use his machinery. If the confectioner had been granted the right to use it, the doctor would have paid him to stop if and only if the doctor's costs were greater than the confectioner's profits.

In other words, private bargaining would ensure that the machinery was used if and only if its use, alongside a payment to compensate the doctor, made both better off. Private bargaining would ensure that its use was Pareto efficient. Bargaining is simply a way to make sure that the confectioner takes account of the private marginal costs of using the machine to produce candy, and also for the external costs imposed on the doctor. That is, the confectioner takes account of the entire social costs. To the confectioner the price of using the annoying machinery (or using it during the doctor's visiting hours) would now send the right message. Private bargaining could be a substitute for liability: it ensures that those harmed would be compensated, and that those who could inflict harm would make efforts to avoid harmful behaviour.

To summarise:

- As long as private bargaining exhausted all the potential mutual gains, the result would (by definition) be Pareto efficient, independently of whether the court upheld Bridgman's right to make noise or Sturges' right to quiet.
- We could object that the court's decision resulted in an unfair distribution of profits, but not that the outcome was Pareto inefficient.

But Coase emphasised that this conclusion was of limited practical relevance because of the costs of bargaining and other impediments to the parties exploiting all possible mutual gains. These costs of bargaining are sometimes called *transaction costs* and in their presence the outcome of bargaining will not be Pareto efficient. If the confectioner cannot find out how badly the noise affects the doctor, the doctor has an incentive to overstate the costs to get a better deal. Establishing each party's actual costs and benefits is also part of the cost of the transaction, and this cost might be too high to make a bargain possible.

Coase also noted that the incentive to compromise depends on who has rights. For example, in one scenario the confectioner might have had to cease production, even though it would have been relatively easy for the doctor to install sound insulation.

Coase's analysis suggests that a lack of established property rights, and other impediments leading to high transaction costs, stop us from using bargaining to resolve externalities. With a clear legal framework in which one side initially owned the rights to produce (or to prevent production of) the externality, as long as these rights were tradable between the two parties in a market for the externality, there might be no need for further intervention.

But recall that bargaining can also fail for other reasons. In Unit 4 players in the ultimatum game sometimes failed to come to an agreement when both parties preferred to walk away empty-handed if the Proposer claimed what the Responder thought was an unfairly large slice of the pie.

10.3 EXTERNAL EFFECTS: POLICIES AND INCOME DISTRIBUTION

Suppose that the government wants to achieve a reduction in the output of bananas to the level that takes into account the costs for the fishermen. (We will assume in this situation that it is not possible to grow bananas without using Weevokil, and that Coasian bargaining is impractical.) There are three other ways this might be done:

- *Regulation* of the amount of bananas produced
- *Taxation* of the production or sale of bananas
- *Enforcing compensation* of the fishermen for the costs imposed on them

Regulation

The government could cap total banana output at 38,000 tonnes, the Pareto-efficient amount. This looks like a straightforward solution. On the other hand, if the plantations differ in size and output it may be difficult to determine and enforce the right quota for each one.

This policy would reduce the costs of pollution for the fishermen, and it would lower the plantations' profits: they would lose their surplus on each tonne of bananas between 38,000 and 80,000.

Taxation

Figure 10.4 shows the MPC and MSC curves again. At the Pareto-efficient quantity, 38,000 tonnes, the MSC is \$400, and the MPC is \$295. The price is \$400. If the government puts a tax on each tonne of bananas produced equal to *\$400 – \$295 = \$105*, the marginal external cost, then the after-tax price received by plantations will be \$295. Now, if plantations maximise their profit, they will choose the point where the after-tax price equals the marginal private cost, and produce 38,000 tonnes, the Pareto-efficient quantity. Use the slideline in Figure 10.4 to see how this policy works.

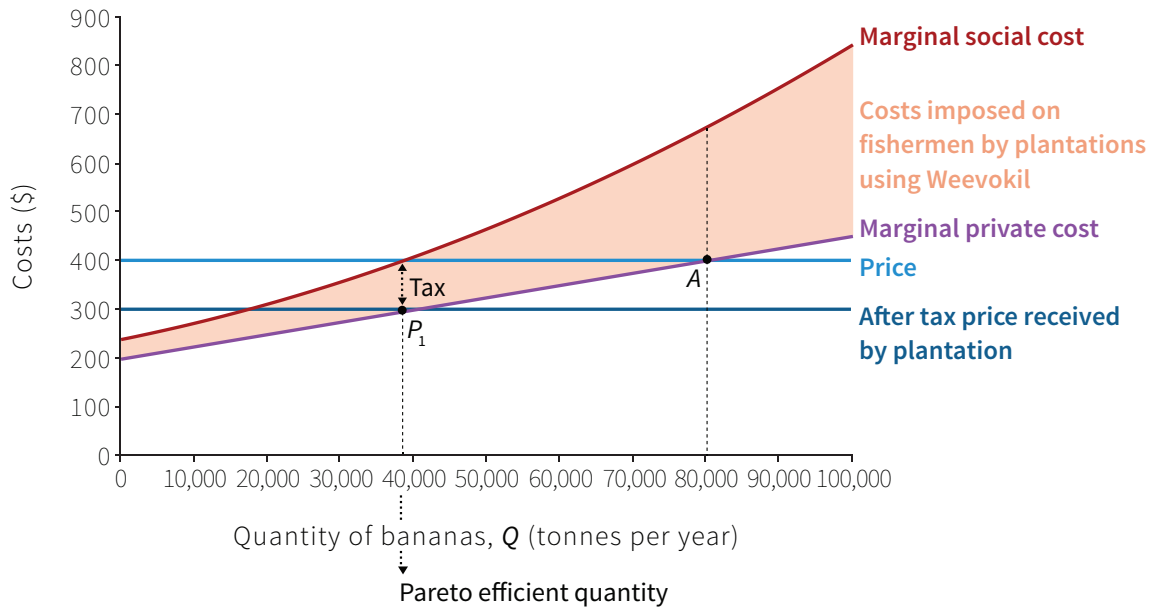


Figure 10.4. *Using a tax to achieve Pareto efficiency.*

The distributional effects of taxation are different from those of regulation. The costs of pollution for fishermen are reduced by the same amount, but the reduction in banana profits is greater, since the plantations pay taxes as well as reducing output; in addition, the government receives tax revenue. The tax corrects the price message, so that the plantations face the full social marginal cost of their decisions. When the plantations are producing 38,000 bananas, the tax is exactly equal to the cost imposed on the fishermen. This approach is known as a *Pigouvian tax*, after the economist who advocated it.

GREAT ECONOMISTS

ARTHUR PIGOU

Arthur Pigou (1877-1959) was one of the first neoclassical economists to focus on welfare economics: the analysis of the allocation of resources in terms of the wellbeing of society as a whole. Pigou won awards during his studies at the University of Cambridge in history, languages and moral sciences (there was no dedicated economics degree at the time). He became a protégé of Alfred Marshall. Pigou was an outgoing and lively person when young, but his experiences as a conscientious objector and ambulance driver during the first world war, as well as anxieties over his own health, turned him into a recluse who hid in his office except for lectures and walks.

Pigou's economic theory was mainly focused on using economics for the good of society, which is why he is sometimes seen as the founder of welfare economics. His book *Wealth and Welfare* was described by Schumpeter as "the greatest venture in labour economics ever undertaken by a man who was primarily a theorist", and provided the foundation for *The Economics of Welfare*. Together, these works built up a relationship between a nation's economy and the welfare of its people. Pigou focused on happiness and wellbeing: he recognised concepts such as political freedom and relative status were important.

Pigou believed that reallocation of resources was necessary when the interests of a private firm or individual diverged from the interests of society, causing what we would today call externalities. He suggested taxation could solve the problem: Pigouvian taxes ensure that producers face the true social costs of their decisions.

Pigou also wrote extensively on the labour side of welfare, such as the link between short-run involuntary unemployment and labour demand, as opposed to the effects of real wages (which he found to be less important than psychological factors).

Despite both being heirs to Marshall's new school of economics, Pigou and Keynes did not see eye-to-eye. Keynes's *The General Theory of Employment, Interest and Money* contained a critique of Pigou's The Theory of Unemployment, and Pigou felt that Keynes's material was becoming too dogmatic and turning students into "identical sausages".

Although overlooked for much of the 20th century, Pigou paved the way for much of labour economics and environmental policy. Pigouvian taxes were largely unrecognised until the 1960s, but they have become a major policy tool for reducing pollution and environmental damage.

Enforcing compensation

The government could require the plantation owners to pay compensation for costs imposed on the fishermen. The compensation required for each tonne of bananas will be equal to the difference between the MSC and the MPC, which is the distance between the red and purple lines in Figure 10.5. Once compensation is included, the marginal cost of each tonne of bananas for the plantations will be the MPC plus the compensation, which is equal to the MSC. So now the plantations will maximise profit by choosing point P_2 in Figure 10.5 and producing 38,000 tonnes. The grey area shows the total compensation paid. The fishermen are fully compensated for pollution, and the plantations' profits are equal to the true social surplus of banana production.

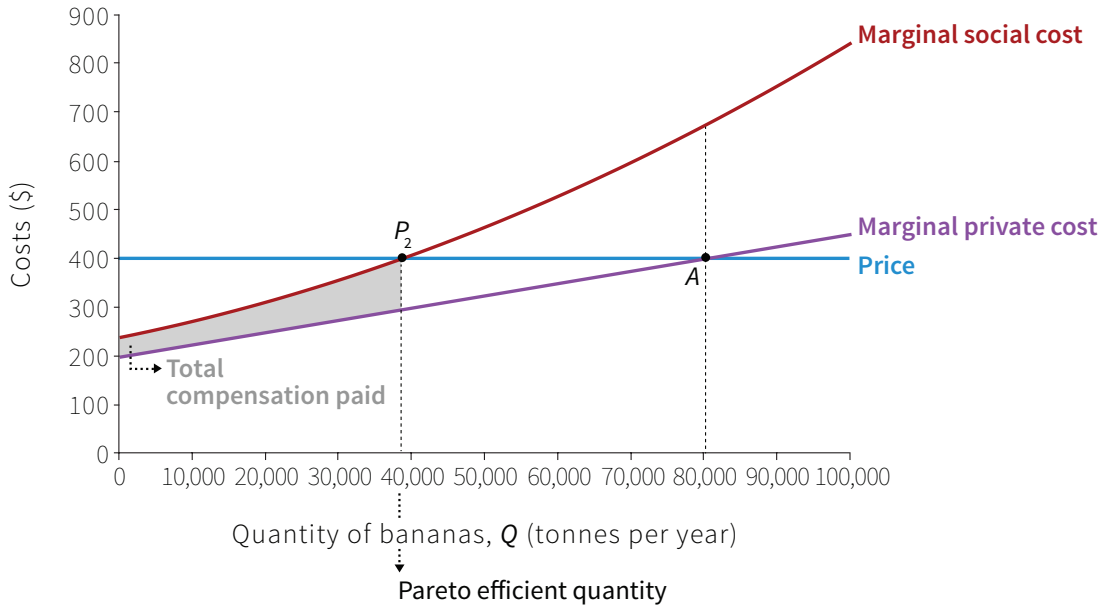


Figure 10.5. *The plantations compensate the fishermen.*

The effect of this policy on the plantations' profits is similar to the effect of the tax, but the fishermen do better—they, rather than the government, receive payment from the plantations.

When we identified 38,000 tonnes as the Pareto-efficient level of output, we assumed that growing bananas inevitably involves Weevokil pollution. So our diagnosis was that too many bananas were being produced, and we looked at policies for reducing production. But that was not the case in Guadaloupe and Martinque—there were alternatives to Chlordecone. If alternatives to Weevokil were available it would be inefficient to restrict output to 38,000 tonnes, because if the plantations could choose a different production method and the corresponding profit-maximising output, they could be better off, and the fishermen no worse off.

So the problem was caused by the use of Chlordecone, not the production of bananas. The market failure occurred because the price of Chlordecone did not incorporate the costs that its use inflicted on the fishermen, and so it sent the wrong message to the firm. Its low price said: "use this chemical, it will save you money and raise profits", but it should have said: "think about the downstream damage, and look for an alternative way to grow bananas".

Of the three policies we considered, requiring the plantations to compensate the fishermen would have given them the incentive to find production methods that caused less pollution and could, in principle, have achieved an efficient outcome. It would also have been better to regulate or tax the sale or the use of Chlordecone rather than the production of bananas, to motivate plantations to find the best alternative to intensive Chlordecone use.

In theory, if the tax on a unit of Chlordecone was equal to its marginal external cost, the price of Chlordecone for the plantations would be equal to its marginal social cost—it would be sending the right message. They could then choose the best production method taking into account the high cost of Chlordecone, which would involve reducing its use or switching to a different pesticide, and determine their profit-maximising output. As with the banana tax, the profits of the plantations and the pollution costs for the fishermen would fall; but the outcome would be better for the plantations, and possibly the fishermen also, if Chlordecone rather than bananas were taxed.

Unfortunately, none of these remedies was used for 20 years in the case of Chlordecone, and the people of Guadaloupe and Martinique are still living with the consequences. In 1993 it was finally recognised the social marginal cost of Chlordecone use was so high that it should be banned altogether.

Like Coasian bargaining, there are limits to how well governments can implement Pigouvian taxes, regulation and compensation, and often for the same reasons:

- The government may not know the degree of harm suffered by each fisherman, so it can't create the best compensation policy.
- While the plantations' private marginal costs are probably well-known, the social marginal costs (including the pollution costs) are more difficult to measure, either to individuals or to society as a whole.
- The government may favour the more powerful group. In this case it could impose a Pareto-efficient outcome that is also unfair.

DISCUSS 10.3: POLICIES

Consider the three policies discussed above. Evaluate the strengths and weaknesses of each policy from the standpoint of Pareto efficiency and fairness.

Now we can add to the table we created in section 10.1. Look at the fifth column, which is new: it adds the possible remedies in the case of negative external effects. Remember, we will combine all our examples of market failure into a large table at the end of this unit, so you will be able to compare possible remedies for other market failures too.

THE DECISION	HOW IT AFFECTS OTHERS	COST OR BENEFIT	MARKET FAILURE (MISALLOCATION OF RESOURCES)	POSSIBLE REMEDIES	TERMS APPLIED TO THIS TYPE OF MARKET FAILURE
A firm uses a pesticide that runs off into waterways	Downstream damage	Private benefit, external cost	Overuse of pesticide and overproduction of crop in which it is used	Taxes, quotas, bans, bargaining, common ownership of all affected assets	Negative external effect, environmental spillovers

10.4 PROPERTY RIGHTS, CONTRACTS AND MARKET FAILURES

In addition to lack of competition as a source of market failure studied in Unit 7, we have now seen another source in the Chlordecone pollution of the Martinique fisheries: external effects. In taking an action so as to maximise profits—choosing the level of banana production or the choice of pesticide—the plantation owners did not take account of the costs imposed on the fishermen. And they had no reason to take account of them: they had the right to pollute the fisheries.

The same is true for the overuse of antibiotics: a self-interested person has no reason to use antibiotics sparingly, because the superbug that may be created will probably infect someone else.

If the price of Chlordecone and the antibiotic was high enough, there would be no overuse. But the price of these goods reflected only the cost to the seller, not the true social cost.

Another example: when fuel costs are low, more people decide to drive to work rather than taking the train. The information conveyed by the low price does not include the environmental costs of deciding to drive. The effects on the decision-maker are termed *private costs and benefits*, while the total effects, including those inflicted or enjoyed by others, are *social costs and benefits*.

Costs inflicted on others (pollution and congestion that are worse because you drive to work) are termed *external diseconomies*; while uncompensated benefits conferred on others are *external economies*.

We can understand why these and other market failures are common by thinking about how they could be avoided.

How could the cost of driving to work accurately reflect all of the costs incurred by anyone, not just the private costs made by the decision maker? The most obvious (if impractical) way would be for the driver to pay everyone affected by the resultant environmental damage (or traffic congestion) an amount exactly equal to the damage inflicted. This is of course impossible to do, but it sets a standard of what has to be done or approximated if the "price of driving to work" is to send the correct message.

Something like this approach applies if you drive recklessly on the way to work, skid off the road, and crash into somebody's house. *Tort law* (the law of damages) in most countries would require you to pay for the damage to the house. You are held liable for the damages so that you would pay the cost you had inflicted on another. Knowing this, you might think twice about driving to work (or at least slow down a bit when you are late). It will change your behaviour and the allocation of resources.

But while tort law in most countries covers some kinds of harm inflicted on others (reckless driving), important external effects would not be covered (adding to air pollution of congestion by driving your car). Here are two further examples:

- *A firm operates an incinerator that produces fumes*: The fumes lower the surrounding air quality. Those being polluted do not have a right to clear air—the right that would be the basis for a claim for compensation from the firm. So the firm does not have to pay these costs.
- *You play music loudly at night and disturb the sleep of the people next door*: Sleeping neighbours do not have an enforceable right not to be woken by your music. There is no way that your neighbours can make you pay them compensation for the inconvenience you cause.

Legal systems also fail to provide compensation for the benefits that one's actions confer on others:

- *A firm trains a worker who quits for a better job*: The trained worker, not the firm, owns the worker's skills. Therefore, even though a different firm receives the benefit, the firm that paid for the training cannot collect compensation from the new firm.
- *Kim, the farmer in Unit 4, contributes to the cost of an irrigation project while other farmers free-ride on Kim's contribution*: Kim has no way of claiming payment for this public-spirited act. The free-riders will not compensate Kim.
- *A country invests in reducing carbon emissions that lowers the risks of climate change for other countries*: As we saw in Unit 4, unless a treaty guarantees compensation for the costs of reduced emission, other countries do not need to pay for this. The environmental improvement for the other countries is an uncompensated benefit.

This failure occurs because the external benefits and costs are not owned by anyone. Think about waste: if you redecorate your house and you tear up the floor or knock down a wall, you own the debris and you have to dispose of it, even if you need to pay someone to take it away. But this is not the case with fumes from the incinerator

or loud music at night. You do not have a contract with the incinerator company specifying at what price you are willing to accept fumes, or with your neighbour about the price of the right to play music after 10pm. In these cases economists say that we have "incomplete, missing, or unenforceable property rights"—or, simply, *incomplete contracts*.

INCOMPLETE CONTRACT

A contract that does not specify, in an enforceable way, every aspect of the exchange that affects the interests of parties to the exchange (or of others).

Incomplete contracts mean that there is no market in which these external effects can be compensated. Therefore economists use the term *missing markets* to describe problems like this.

In the case of Weevokil pollution:

- *The fishermen's property rights were incomplete*: They did not include the right not to have fisheries polluted.
- *Any contract between the two parties must be incomplete*: It could not cover the damage done by each plantation's use of the pesticide on each fisherman's livelihood.

Why don't countries just rewrite their laws so that benefits conferred on others must be rewarded, and costs inflicted on others be paid by the decision-maker?

In most cases it is impractical to use tort law to make people liable for the costs they inflict on others, because we don't have that information. But it is equally infeasible to use the legal system to compensate people for the beneficial effects they have on others, for example, to pay those who keep beautiful gardens an amount equal to the pleasure this confers on those who pass their house, because a court would have to know how much that pleasure was worth to each passerby.

To understand how this information problem, we introduce these new concepts: verifiable and asymmetric information.

VERIFIABLE INFORMATION, ASYMMETRIC INFORMATION

- Information that can be used to enforce a contract is *verifiable*.
- Information that is known by one party but not by others is *asymmetric*.

Recall from Unit 6 that an economic interaction is said to be governed by a *complete contract* if the contract covers all of the aspects of the interaction that are of interest to anyone affected, and the contract can be enforced at zero (or minimal) cost by any of the parties. We studied the employment contract, which is *incomplete* because it covers the employee's time but not the employee's level of effort, even though employee effort affects the employer's profits. Because the contract is incomplete, a hard-working employee is conferring a benefit on her employer without being compensated for the work done.

The labour contract is incomplete partly because the information about how hard she worked is asymmetric: the employee knows if she could work harder, but the boss does not.

And even if the boss did know, his information is typically not verifiable; he cannot reclaim her wages just by reporting to a court that he caught her doing private email.

In the five bulleted examples earlier in this section, the reason why uncompensated external costs and benefits occur is the same:

- Some information that is of concern to someone other than the decision-maker is asymmetric or non-verifiable.
- Therefore there can be no contract or property rights ensuring that external effects will be compensated.
- As a result, some of the social costs or benefits of the decision-maker's actions will not be included (or will not be sufficiently important) in the decision-making process.

DISCUSS 10.4: INCOMPLETE CONTRACTS

In each of the five cases (incinerator, loud music, training, irrigation and climate change) above:

1. Explain why the external effects are not (and possibly could not be) covered by a complete contract.
2. How this can be traced to the fact that some critical information that the contract would require is asymmetric or non-verifiable (explain which critical piece of information has this characteristic)?

10.5 PUBLIC GOODS

Markets are not the best way to determine the allocation of some kinds of goods or services. We saw in Unit 6 that, within firms, tasks are assigned and resources allocated by the command of management, not by the working of supply and demand. As Coase pointed out, a reason firms exist is that markets are not always the least-cost way to allocate resources. There is another large class of goods and services for which this is also true. These are called *public goods* and they include such things

as a system of justice, national defence and weather forecasting, services that are typically provided by governments rather than the market. Other examples are the knowledge of the rules of multiplication, or the view of the setting sun.

The defining characteristic of a public good is that if it is available to one person it can be available to everyone at no additional cost. For a view of the setting sun, one more person enjoying it does not deprive anyone else of enjoyment. This means that, once the good is available at all, the marginal cost of making it available to additional people is zero. Goods with this characteristic are sometimes called *non-rival goods*.

Pure public goods are non-rival goods from which others cannot be excluded. Examples include a view of a lunar eclipse, knowing the time of day, and publicly broadcast signals, such as weather forecasts or the news, for people in a particular area.

PUBLIC AND NON-RIVAL GOODS

A good is *public* if:

- Its use by one person does not reduce its availability to others.
- Note that a good that, if available to anyone, is available to everyone at no additional cost is called non-rival.

For some public goods it is possible to exclude additional users, even though the cost of their use is zero. Examples are satellite TV, the information in a copyrighted book, or a film shown in an uncrowded cinema; it costs no more if an additional viewer is there, but the owner can nonetheless require than anyone who wants to see the film must pay. The same goes for a quiet road on which tollgates have been erected. Drivers can be excluded (unless they pay the toll) even though the marginal cost of an additional traveller is zero. Public goods from which people may be excluded are sometimes called *artificially scarce goods* or club goods. ("Club" because they function like joining a private club: when the golf course is not crowded, adding one more member costs the golf club nothing, but the club will still charge a membership fee.)

The opposite of public goods are *private goods*. Like the loaves of bread, dinners in restaurants, rupees divided between Anil and Bala in Unit 4, and boxes of breakfast cereal that we have used as examples so far, private goods are both *rival* (more for Anil means less for Bala) and excludable (Anil can prevent Bala from taking his money).

There is a fourth kind of good that is rival, but not excludable, called a *common-pool resource*. Examples include fisheries open to all: what one fisherman catches cannot be caught by anyone else, and anyone who wants to fish can do so. Figure 10.6 summarises the four kinds of goods.

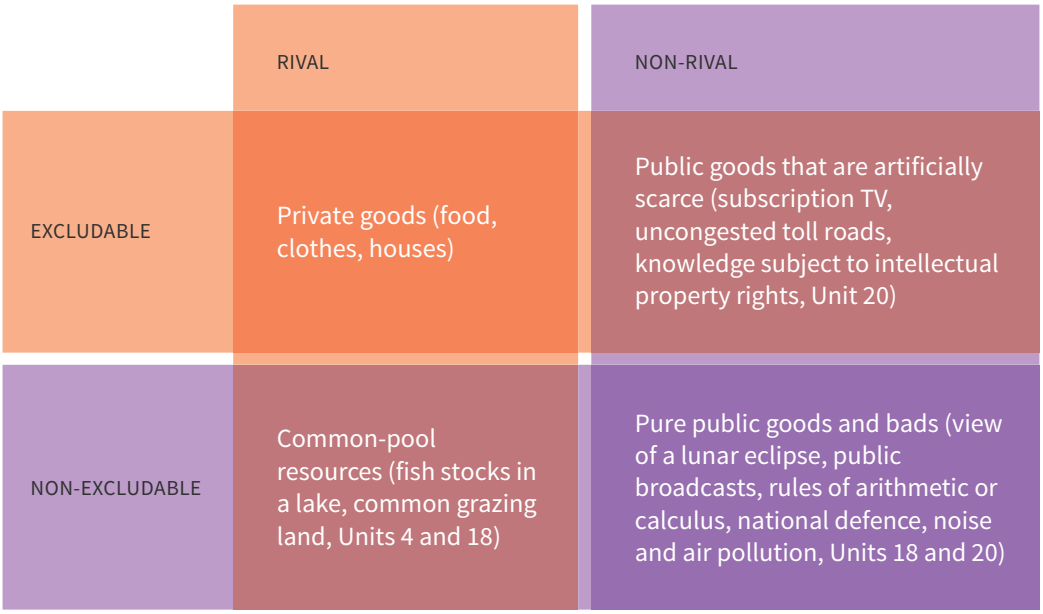


Figure 10.6 *Private goods and public goods.*

As can be seen from the examples, whether a good is private or public depends not only on the nature of the good itself, but on legal and other institutions:

- Knowledge that is not subject to copyright or other intellectual property rights would be classified as a *pure public good*…
- … But when the author uses copyright law to create a monopoly on the right to reproduce that knowledge, it is a *public good that is artificially scarce*.
- Common grazing land is a *common-pool resource*…
- … But if the same land is fenced to exclude other users, it becomes a *private good*.

Markets typically allocate private goods. But, for the other three kinds of good, markets are either not possible or likely to fail. There are two reasons:

- *When goods are non-rival the marginal cost is zero*: Setting a price equal to marginal cost (as is necessary for a Pareto-efficient market transaction) will not be possible unless the provider is subsidised.
- *There is no way to charge a price for the good or service*: The provider cannot exclude additional users.

Therefore, when goods are not private, it is not easy for governments to create public policy to achieve both Pareto-efficient and fair outcomes. In the next section we examine knowledge and intellectual property rights, and in Unit 18 we return to common-pool resources and public bads.

DISCUSS 10.5: RIVALRY AND EXCLUDABILITY

For each of the following goods or bads, decide whether they are rival and whether they are excludable, and explain your answer. If you think the answer depends on factors not specified here, explain how.

1. A public lecture at a university
2. Noise produced by aircraft around an international airport
3. A public park
4. A forest used by local people to collect firewood
5. Seats in a theatre
6. Bicycles available for hire to the public to travel around a city

In our table we can show how external effects that are not compensated produce market failures:

THE DECISION	HOW IT AFFECTS OTHERS	COST OR BENEFIT	MARKET FAILURE (MISALLOCATION OF RESOURCES)	POSSIBLE REMEDIES	TERMS APPLIED TO THIS TYPE OF MARKET FAILURE
You take an international flight	Increase in global carbon emissions	Private benefit, external cost	Overuse of air travel	Taxes, quotas	Public bad, negative external effect
You travel to work by car	Congestion for other road users	Private cost, external cost	Overuse of cars	Tolls, quotas, subsidised public transport	Common pool resource, negative external effect
A firm invests in R&D	Other firms can exploit the innovation	Private cost, external benefit	Too little R&D	Publicly funded research, subsidies for R&D, patents	Public good, positive external effect

10.6 INNOVATION AND DIFFUSION

The television programme *Dragons' Den* gives inventors three minutes to pitch an idea for a new product to potential investors. They hope the investors will back them with the finance needed to set up in business. We know more about the successes, such as Reggae Reggae Sauce (created in 2007 by an entrepreneur called Levi Roots, real name Keith Valentine Graham), and the ideas that had no hope of being funded (a glove for drivers to wear on one hand to remind them which side of the road to drive on) than about the credible investments that later failed.

In Unit 2 we learned that innovations leading to new products (product innovations like Toyota's hybrid car) and reduced costs of producing existing products (process innovations, like the spinning jenny) were an essential part of the capitalist revolution that dramatically raised living standards in the countries where it happened.

For consumers and other firms, this process of copying an innovation reduces prices and makes available desirable commodities. Since the 1980s competition between the world's top mobile phone companies such as Samsung (South Korea), Nokia (Finland), and Apple (US) has stimulated a continuous process of product innovation and improvement in design and capabilities, and further price competition. In 1996 Nokia combined a mobile phone and Personal Digital Assistant in a single device, and the smartphone was born. Other companies—Ericsson, Palm, Blackberry and NTT Docomo—developed the idea further, followed by Apple's touchscreen device, the iPhone, in 2007.

In Unit 2 we also explained how competition among firms propelled this process, providing both:

- *The carrot*: The desire to capture innovation rents, that is, profits above the opportunity cost of capital
- *The stick*: The fear in those who lagged in the innovation process of being undersold or outcompeted by early innovators

Here we ask how well this process of competition for innovation rents (or merely survival) works, and whether it could be made to work better. The carrot-and-stick interpretation of the permanent technological revolution underscores two necessary aspects of the process:

- *Innovation*: The development of new products and new processes by an individual or firm
- *Diffusion*: The use of the novel methods or a new product by a large number of users

Ideally a combination of market competition and government policy would promote both innovation and diffusion. But this, we will see, is not easy to do. The reason is that such a combination would have to accomplish two objectives that may not be consistent:

- *Innovation*: Policy needs to provide incentives to develop new products and processes
- *Diffusion*: Policy should make it easy for large numbers of users to use these methods

There may be a trade-off between the two objectives for the following reasons:

- Providing incentives for those who make costly investments in the innovation process implies that they receive adequate innovation rents when they are successful.
- Therefore they must, at least temporarily, be able to charge a price above the average cost of producing the good (including the opportunity cost of capital).
- This can occur if the innovator can keep the details of the innovation secret, or if the government grants the innovator a patent or copyright that, for a fixed period, prevents others from copying the idea.
- *But this restricts diffusion*: Other firms (and many customers) are deprived of, or have delayed access to, the benefits of innovation.

A patent is a right of exclusive ownership of an idea, which lasts for a specified length of time (typically 20 years). When the innovation policy is the granting of patents or copyrights, the conflict between innovation and diffusion arises for a simple reason, easily understood in terms of the model of the firm in Unit 7: patents and copyrights work because they create an artificial monopoly of the use of some novel idea, whose use by others would be beneficial for consumers and other firms alike. An extreme example of the beneficial effect of diffusion is the vaccine against poliovirus, created by Jonas Salk. When he was asked who owned the patent, he replied: “The people, I would say. There is no patent. Could you patent the sun?”

Patents convert a public good (knowledge of the new process or product) into a private good that can be owned, and from which others can be excluded. Once a new product or process is known, others can profit from it, free-riding on the original investment. Governments can address free-riding by granting a patent to the innovator. There are many examples of profitable use of patents: the prolific inventor Thomas Edison (1847-1931), for example, made a fortune from his innovations in telegraphy and electric power distribution.

Innovation involves a delicate balancing act for government policy. To highlight the trade-off faced by the policymaker, we take the case of a successful cost-reducing process innovation in the bread industry studied in Units 8 and 9, where the entrepreneur continues to set the market price. With a lower cost of production, the

innovator earns an economic rent on each unit of output sold. In Figure 10.7a, the entrepreneur takes a forward-looking approach to the decision to incur the costs of a risky innovation:

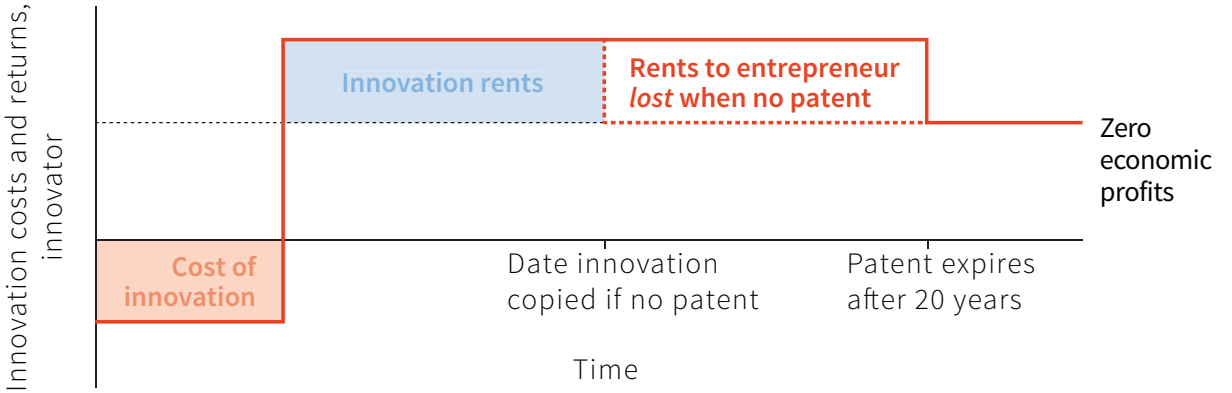


Figure 10.7a *Costs and returns from innovation to the innovator.*

Following a successful innovation, even if there is no patent protecting the innovator from copying, the entrepreneur earns innovation rents because it takes time for followers to copy and achieve the lower costs of the innovator. Note that once the lower-cost method spreads across the industry, the price of bread falls as analysed in Discuss 9.1 and profits in the industry return to normal. If, however, the innovation is protected by a patent, innovation rents are earned for longer—until the patent expires after 20 years. This example illustrates that from the perspective of the entrepreneur, patent protection increases the rewards from successful innovation and therefore boosts the incentive to innovate.

Until the innovation is copied, the entrepreneur is the only beneficiary since the price of bread is unchanged. Figure 10.7b show the benefits from diffusion:

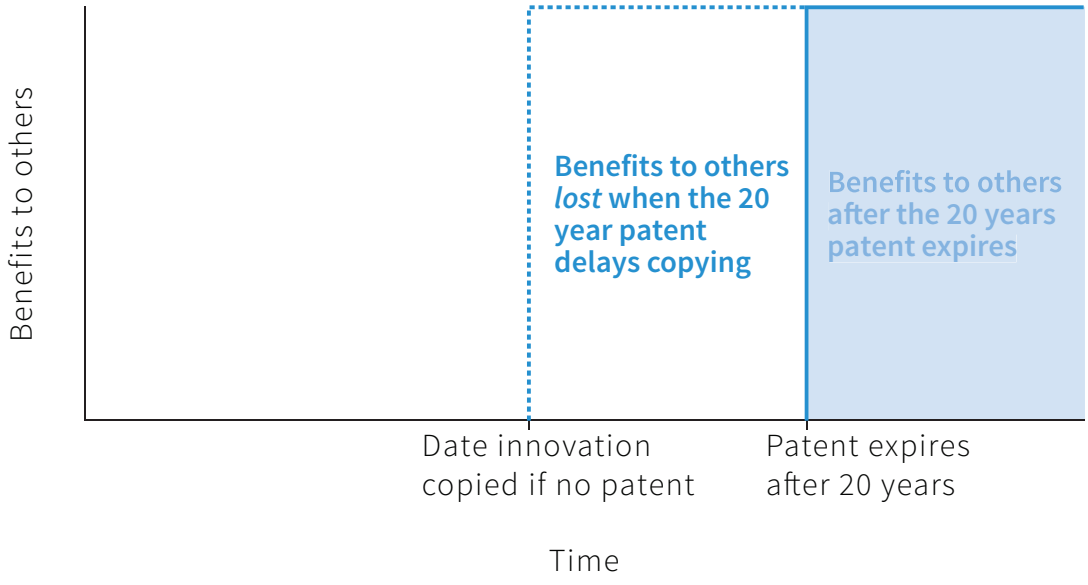


Figure 10.7b *Benefits from diffusion of an innovation.*

The dilemma for the design of patent policy is that making it easier to copy innovations may initially increase the benefit to others from diffusion, as in Figure 10.7b. But it reduces the size of the shaded area above the zero-economic-profits line in Figure 10.7a. This, in turn, reduces the incentive for entrepreneurs to innovate, which reduces the benefits to all of us if there is less innovation. On the other hand, many believe that the existing patent system offers too much protection for inventors, which restricts diffusion, as these two articles explain.

Our model of the innovating firm suggests two other types of policy that might ease the tension between diffusion and innovation:

- Polices that *reduce the cost of innovation*
- Policies that *provide greater innovation rents* without giving the innovator a monopoly

Reducing the cost of innovation

Policies reduce the size of the "cost of innovation" rectangle in Figure 10.7a would mean entrepreneurs could justify a decision to innovate, even if innovation rents were smaller. Policies that reduce the cost of innovation include governmental support for science education at all levels, basic research for which results will not be patented, and direct subsidies to innovating firms. As a result, a firm might invest in innovation even if there was a shorter patent protection period.

Publicly supported education and research reduce the cost to a firm of innovation because they create an abundance of engineers and new ideas. Finally, research subsidies to particular firms or industries are targeted in ways that may reflect government or other priorities: for example, reflecting a need to develop low-cost renewable energy sources or effective drugs for the treatment of a particular disease.

Increasing innovation rents without granting a monopoly

A different type of policy may be effective in cases where the need for a particular innovation is well-defined. Imagine a competition offering a cash prize to the first innovation satisfying a particular need: for example, the first firm to create a low-cost, easy-to-administer drug to treat sleeping sickness (a major killer in Africa). The prize would be the innovation rent for the inventor, and the need for a patent disappears.

10.7 POSITIONAL EXTERNAL EFFECTS

Some goods, such as cars and clothes, may act as status symbols. Their owners value them partly because they rank them above other people.

Perhaps one of your motives when you buy a car, or a coat, is to demonstrate your wealth and superior style. Or perhaps you settle for a cheaper second-hand coat but feel envious, or embarrassed, or disadvantaged at a job interview. Thorstein Veblen (1857-1929), an economist and sociologist, described buying luxury items as a public display of social and economic status as *conspicuous consumption*.

Goods that are valued more because they are expensive are an example of a larger class called *positional goods*. They are positional because they are based on status or power, which can be ranked as high or low. Our positions in this ranking, like the rungs of a ladder, may be higher or lower. But there is only a fixed amount of a positional good to go around. If Jo is on a higher rung of the ladder because of her new coat, somebody must now be on a lower rung.

The effect of positional goods on other people is a negative external effect. To see its implications, consider the case of Sue Smith and her sister Jo Jones, each moving with their families to a new town. Each family has a choice between buying a luxury house or a more modest one. Their payoffs are represented in Figure 10.8. Since both families have limited funds they would be better off if both bought modest houses than if both bought luxury ones, squeezing the rest of their budget. But Sue is status-conscious, and the two families are competitive when it comes to lifestyle: if the Smiths buy a modest house, the Joneses can benefit from feeling superior if they choose a luxury house. The Smiths, in turn, will feel miserable.

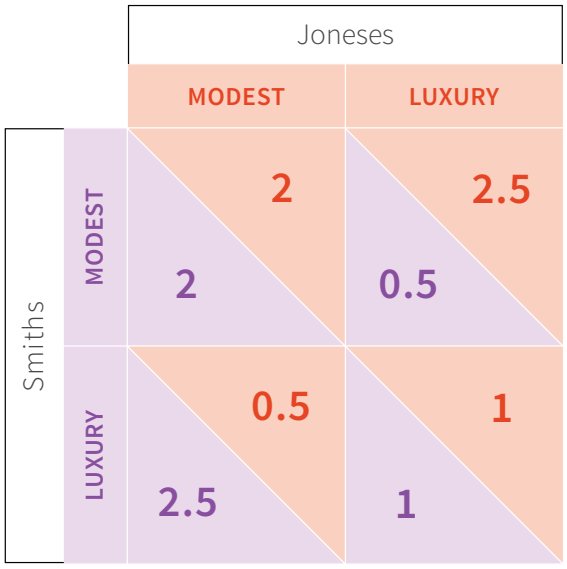


Figure 10.8 *Keeping up with the Joneses.*

You can see that this problem has the structure of a *prisoners' dilemma*. Whatever the Joneses do, the Smiths are better off with a luxury house. For both couples, choosing Luxury is a dominant strategy. They will achieve a payoff of 1 each, and the outcome is Pareto inefficient, because both would be better off if they bought modest houses. The root of this problem is the external cost that one family imposes on the other

by choosing a luxury house. The price of the luxury house that Sue's family will buy does not include the positional externalities that purchasing it inflicts on her sister's family. If it did, Sue would not buy the luxury house, given the payoffs in the table.

What can they do to avoid the Pareto-inefficient outcome? We know from Unit 4 that altruism would help, but these families are not altruistic about houses. Following Coase's advice, they could agree in advance that if one family has a better house they will compensate the other, with a payment chosen to make Modest a dominant strategy. However, courts might be unwilling to enforce a contract like this.

The "keeping up with the Joneses" problem that Sue and Jo face arises because people care not only about what they have, but also about what they have *relative* to what other people have. This is sometimes called a *Veblen effect*.

Veblen effects help to explain two facts about modern economies:

- *People work longer hours in countries in which the very rich receive a larger fraction of the income*: For example, the US has both higher hours of work and a higher income share of the very rich than Germany, France, Sweden and the Netherlands. The rich are the "Joneses" who people want to keep up with. To do this, they work longer hours if the Joneses are richer. A century ago American workers worked fewer hours than workers in all these countries. But over the past 100 years the share of income going to the very rich declined in all these countries. Sweden, for example, went from one of the most unequal countries (by this measure) to one of the more equal. (For more on this topic, watch Juliet Schor's *Economist in Action* video in Unit 3)
- *As a nation gets richer, its citizens often do not become happier*: When an individual gets a wage rise or loses a job, it has a large effect on how happy that individual claims to be. But economists have also found that a change in our income has a much smaller effect if most of our acquaintances also got the same rise, or also lost their job. When an entire nation gets richer, the effect on individual happiness is small, if there is one at all.

This is a Veblen effect—just as Sue being happier with a modest house if her sister has one too. When Veblen effects are present the conspicuous consumption of the well-off is a positional good and a negative external effect: if it is experienced by everyone, reducing their satisfaction with their own situation, it is a public bad.

DISCUSS 10.6: VEBLEN EFFECTS AND POLICY

1. The example of Veblen effects above concerns houses. Can you think of any other examples where Veblen effects are present?
2. Why do Veblen effects cause inefficiency?
3. Describe in what way Veblen effects are similar to (or different from) pollution?
4. Discuss whether the government should adopt policies to address this market failure and, if so, what might they be?

10.8 MISSING MARKETS: INSURANCE AND LEMONS

The functioning of markets is affected when one person knows something relevant that the other person doesn't know. This is called *asymmetric information* and takes two forms, both of which affect how markets work. We have already studied one asymmetric information problem in detail in the labour market in Unit 6. That is a case called *hidden actions*. In this section, we introduce the other problem, that of *hidden attributes*. We will see that insurance markets are characterised by problems of hidden actions and hidden attributes:

- *Hidden actions*: An employee knows how hard she is working. The employer does not. Some of the employee's actions are *hidden* from the employer. This creates a problem called *moral hazard*. Because the employer cannot observe the worker's effort accurately, the employer creates an incentive for effort by paying a wage above the reservation wage. More effective monitoring of worker effort would reduce the information asymmetry and the extent of market failure.
- *Hidden attributes*: When you want to purchase a used car, for example, the seller knows the quality of the vehicle. You do not. This attribute of the car is *hidden* from the prospective buyer. The problem caused by hidden attributes is called *adverse selection*.

Hidden attributes and adverse selection

A famous example of how hidden attributes may result in a market failure is known as the market for lemons. A "lemon" is slang for a used car that you discover to be defective after you buy it. The market for lemons describes a model of a used car market:

- Every day, 10 owners of 10 used cars consider selling.
- The cars differ in quality—which we measure by the true value of the car to its owner. Quality ranges from zero to $9,000 in equal steps: there is one worthless car, one worth $1,000, another worth $2,000, and so on. The average value of the cars is thus $4,500.
- There are many prospective buyers and each would happily buy a car for a price equal to its true value, but not more.
- Sellers do not expect to receive the full value of their vehicle, but they are willing to sell if they can get more than half the true value. So the sum of the surplus—buyers' and sellers'—will be half the price of the car.

If prospective buyers approach each seller and bargain over the price, by the end of the day all of the cars (except for the entirely worthless one) would be sold at a price somewhere between their true value and half the true value. The market would have assured that all mutually beneficial trades would take place.

But, on any day, there is a problem: potential buyers have no information about the quality of any car that is for sale. All they know is the true value of the cars sold the previous day. The most that prospective buyers are willing to pay for a car will be the average value of the cars sold the day before.

Now suppose that 10 cars had been offered on the market the day before. We use proof by contradiction to show that, one by one, the highest quality cars will drop out of the market, until there is no market in used cars. Consider the market today:

- Yesterday all the cars (as we assumed at the start) were put on the market and sold.
- The average value of these cars was $4,500, so the most a buyer is willing to pay today will be $4,500.
- At the beginning of the day, each prospective seller considers selling his or her car, expecting a price of $4,500 at the most. Most of the owners are happy: it is more than half the true value of their car.
- But one owner isn't pleased. The owner of the best car would not sell unless the price *exceeds* half the value of his car: more than $4,500.
- Prospective buyers will not pay this price. So today the owner of the best car will not offer it for sale. No one with a car worth $9,000 will be willing to participate in this market.
- The rest of the cars will sell today: their value averages $4,000.
- Tomorrow buyers will know the average value of the cars sold today. And so tomorrow, buyers will decide they will be willing to pay at most $4,000 for a car.
- The owner of tomorrow's highest-quality car (the one worth $8,000) will know this, and that she will not get her minimum price, which is greater than $4,000. Tomorrow, she will not offer her car for sale.

- As a result, the average quality of cars sold on the market tomorrow will be $3,500, which means the owner of the third-best car will not put his car up for sale the day after tomorrow.
- And so it goes on, until, at some point next week, only the owner of a lemon worth $1,000 and a totally worthless car will remain in that day's market.
- If cars of these two values had sold the previous day, then, the next day, buyers will be willing to pay at most $500 for a car.
- Knowing this, the owner of the car worth $1,000 will decide she would rather keep her car.
- The only car on the market will be worth nothing: cars that remain in this market are lemons, because only the owner of a worthless car would be prepared to offer that car for sale.

Economists call processes like this *adverse selection* because the prevailing price selects which cars will be left in the market. They will definitely be of lower quality, and therefore adverse, from the point of view of the buyer.

DISCUSS 10.7: HIDDEN ATTRIBUTES

Consider the following markets in which hidden attributes may be an impediment to market participants being able to exploit all of the possible mutual gains from exchange:

1. A second-hand good being sold on eBay, Craigslist or a similar online platform
2. Renting apartments through AirBnB
3. Restaurants of varying quality

Explain how the following may facilitate mutually beneficial exchanges, even in the presence of hidden attributes:

4. Electronic ratings shared among past and prospective buyers and sellers
5. Exchanges among friends, and friends of friends
6. Trust and social preferences
7. Intermediate buyers and sellers, such as used car dealers

Adverse selection and moral hazard in the market for insurance

The market for lemons is a well-known term in economics, but the lemons problem—that is, the problem of hidden attributes—is not restricted to the used car market.

To see why, think about health insurance. Imagine hypothetically that you will be born into a population, but do not know whether you would be born with a serious health problem, or might contract such a problem later in life, or perhaps be entirely healthy until old age. Would you buy health insurance if the cost of the premium (which is the same for everyone) would be sufficient to pay for the medical services you would need, if everyone were to sign up?

In this situation most people would be happy to purchase the health insurance described above, because serious illness imposes high costs that are often impossible for an average family to pay. The costs of protecting you and your family from a financial catastrophe (or the possibility that you can't afford healthcare when you need it) are worth the insurance premium.

The thought experiment is unrealistic: we cannot ask a question "before people know how healthy they will be", because that means asking them before they are born. It is another use of John Rawls' veil of ignorance that we discussed in Unit 5. It helps us to think about a problem as an impartial observer.

In this case it makes an important point. Though everyone would have bought insurance if they did not know about their future health status, the situation changes dramatically if we can choose whether to buy health insurance *without the veil of ignorance*, that is, knowing our health status. Look at the situation from the standpoint of the insurance company:

- People are more likely to purchase insurance if they know that they are ill. So the average health of people buying insurance will be lower than the average health of the population.
- *This information is asymmetric*: The person buying the insurance knows how healthy he or she is, but the insurance company does not.
- Therefore insurance companies will be profitable only if they charge higher prices than they would charge if all members of the population were forced to purchase the same insurance.
- In which case, the price will be high enough that only people who knew they were seriously ill would wish to purchase insurance. This is a case of adverse selection.
- So, to remain in business, the insurance companies will now have to charge even higher prices. Eventually the vast majority of the people purchasing insurance will be those who know they already have a serious health problem.
- Healthy people who want to buy insurance in case they fall ill in the future are priced out of the market, and will not buy insurance.

In this case we have another example of a *missing market*, this time for health insurance. It is a market that could exist, but only if health information were symmetrical and verifiable (ignoring for the moment the problem of whether everyone would want to share their health data). It could provide benefits to both insurance company owners and people who wanted to insure themselves. Not having such a market is Pareto inefficient.

To address the problem of adverse selection due to asymmetric information, and the resulting missing markets for health insurance, many countries have adopted policies of compulsory enrolment in private insurance programs or universal tax-financed coverage.

Hidden attributes are not the only problem facing insurers, whether private or governmental. There is also the problem of hidden actions: buying the insurance policy may make the buyer more likely to take exactly the risks that are now insured. For example, a person who has purchased full coverage for his car against damage or theft may take less care in driving or locking the car than someone who had not purchased insurance.

Insurers typically place limits on the insurance they sell. For example, coverage may not apply (or may be more expensive) if someone other than the insured is driving, or if it is usually parked in a place where a lot of cars are stolen. These provisions can be written into an insurance contract.

But the insurer cannot enforce a contract about how fast you drive or whether you drive after having had a drink. These are the actions that are hidden from the insurer because of the asymmetric information: you know these facts, but the insurance company does not.

As in the case of labour effort, hidden actions in these cases lead to a market failure because the private benefit to the driver (driving as fast as she prefers) may end up inflicting an external cost on the insurance company. This section of our table sets out the problems of, and possible remedies for, hidden actions and attributes.

THE DECISION	HOW IT AFFECTS OTHERS	COST OR BENEFIT	MARKET FAILURE (MISALLOCATION OF RESOURCES)	POSSIBLE REMEDIES	TERMS APPLIED TO THIS TYPE OF MARKET FAILURE
An employee on a fixed wage decides how hard to work	Hard work increases her employer's profits	Private cost, external benefit	More effort and higher wages would be better for both worker and employer	More effective monitoring to make the contract more complete, reduced conflict of interest between employer and worker	Incomplete labour contract (does not cover effort), hidden action, moral hazard
An unemployed worker offers to work as hard as an employed worker at a lower wage	Would confer a benefit on the employer	External benefit	Despite this being a mutually beneficial agreement, it could not be enforced so the employer refuses, involuntary unemployment	More effective monitoring to make the contract more complete, reduced conflict of interest between employer and worker	Incomplete labour contract (does not cover effort), hidden action, moral hazard
A person who knows he has a serious health problem decides to purchase insurance	Results in insurance provider making losses	Private benefit, external cost	Those whose higher risk exposure is known only to them (not the insurance company) will be more likely to buy	Mandatory purchase of health insurance, public provision, mandatory health information sharing	Missing markets (hidden attribute, adverse selection)
A person who has purchased car insurance takes more risks	More prudent driving would contribute to insurance company profits	Private benefit to the insured, external cost for the insurance company	Insurance is more expensive than it would be were there no hidden actions, too little insurance purchased	Installing monitoring devices that generate verifiable information on driving habits	Missing markets (hidden action, moral hazard)

10.9 DO MARKETS CAUSE INEQUALITY?

When we analyse the distributional consequences of an individual market, we measure what individuals gain as a result of participating—their surplus. We have seen, for example, that a firm with market power can increase its own surplus, and reduce that of consumers, by setting a high price. But what determines inequality in the distribution of resources in a market economy as a whole?

Do markets cause inequality? In one situation we can certainly answer "no". This is the equilibrium of a market like the one we studied in Unit 8, *in which all buyers and sellers are price-takers, and the market clears in equilibrium.*

To see why this is so, imagine the many islands making up a hypothetical country, which we call *Walrasia*. We named it after Leon Walras, the Great Economist from Unit 8 who was the architect of the theory of competitive equilibrium in a model with price-taking buyers and sellers, just like the one that prevails in Walrasia. On each island the natural resources and the skills of the population are suitable for only one kind of economic activity. The residents of Wheat Island grow grain and make bread, Goat Island produces milk and meat, Coal Island is populated by miners, on Cotton Island people grow cotton and make clothes.

Walrasia has no firms; on each island, each family owns an equal amount of the land and other resources necessary for its work. Residents of all islands work equally hard, but they lack the skills and are not able to acquire the resources necessary to go into any other line of work except the one in which their island specialises. At the weekly market (on Market Island, of course), large numbers of buyers and sellers from all over the country buy and sell their produce.

In the equilibrium of the markets all buyers and sellers will act as price-takers and the markets will clear. A consequence is that the *law of one price* will operate, meaning that everyone transacts at the same prices.

But the similarities end there. The distribution of income in Walrasia is very unequal. There is a small island that produces chocolate, which is in high demand all over Walrasia. Since there are relatively few suppliers, the equilibrium price is high and the chocolate producers enjoy a high standard of living. Miners are poor, however; although they are skilled workers, coal is abundant, and an alternative source of energy is available from Oil Island. So the price of coal is low.

Why are some Walrasians rich and others poor? They all work equally hard, so this seems unfair. Our first instinct is to blame the market for the inequality we see in Walrasia: after all, Walrasians bought and sold their goods on a market, and they ended up with very different levels of income.

But the market has not created this inequality. Whether a Walrasian is rich or poor is entirely determined by where that Walrasian starts: in this case, on which island he or she is born. We will see in Unit 19 that economic inequalities in the world today share this feature with Walrasia: most disparities can be traced to who your parents are, including in which country you were born. In Walrasia:

- *People with the same endowments (that is, residents of the same island) end up with the same incomes after trading*: The residents of each island have the same skills and resources, and they all buy and sell at the same prices.
- *People with different endowments (residents of different islands) end up with incomes that differ to exactly the same extent that the islands differ in the value of their endowments.*

Some families were born on islands with the resources and skills to produce highly valued goods; others have endowments that are worth far less, because people don't want to buy what they produce, keeping demand low, or because there are many others with similar endowments, maintaining a plentiful supply. The prices that equate supply and demand in Walrasia mean that the endowments of some are worth much more than others. The difference in living standards among the Walrasians after they have exchanged their goods are exactly the same as the differences in their endowments before they trade. For an example of how trade does not alter the level of inequality in Walrasia, read this Unit's Einstein section.

But Walrasia is a hypothetical place: everyone is a price-taker and the law of one price always holds. Think of markets that are closer to the real-world conditions we have studied, in which this does not apply:

- *Bargaining power*: Walrasia's economy is just as before, but now when Cotton Islanders go to Market Island they are not price-takers. They form a single organisation and set the price for cotton goods that will maximise their profits. In this case the non-competitive market shifts the distribution of income in favour of the Cotton Islanders.
- *Discrimination*: Walrasia's population is now made up of two ethnic groups, the greens and the blues, where the greens greatly outnumber the blues on each island. Greens discriminate against blues: they charge higher prices to blues. This violates the law of one price, and this market contributes to greater inequality. Inhabitants of an island who were born with identical endowments will not all have identical incomes.
- *Unemployment*: We leave Walrasia and visit another country comprising a set of islands where, unlike Walrasia, production is organised in firms. Employers can employ workers from other islands, and have to pay a wage above the reservation wage in order to get workers to work hard. On Labour Market Island we find that, among the identical workers arriving hoping for a job, some return to their islands unemployed and without a wage. People who are otherwise identical to them find jobs, and are paid a wage. Once again, the law of one price has failed. The price at which the employed workers sell their time is the wage, but the unemployed are

unable to sell their time at any price. Workers with identical endowments now have different incomes. Changing how this market works—lowering the level of unemployment—would result in a more equal distribution of income.

DISCUSS 10.8: POLICYMAKING IN WALRASIA

Consider the initial case of the islands of Walrasia (before bargaining power and discrimination were introduced). How would you evaluate the following policies?

1. A tax on chocolate and wheat, with the proceeds used to subsidise the incomes of the miners on Coal Island.
2. A redistribution of land from each family of Chocolate Island and Wheat Island to some of the miners (assume they would also receive training to make chocolate and grow wheat).
3. Doing nothing.

10.10 THE LIMITS OF MARKETS

Markets might seem to be everywhere in the economy, but this is not the case. Recall Herbert Simon's image from Unit 6 of a Martian viewing the economy who sees green fields, which are firms connected by red lines of buying and selling in markets. Families do not allocate resources among parents and children by buying and selling. Governments use the political process rather than market competition to determine where, and by whom, schools will be built and roads maintained.

Why are some goods and services allocated in markets, while firms, families and governments allocate others? Why not just put everything up for sale?

People disagree about the appropriate extent of the market, some thinking that some things that are now for sale should be allocated by other means, while others think that markets should take a larger role in the economy.

Those who wish to limit the extent of the market often make two arguments:

- *Other institutions may be more effective*: Firms, families, other private bodies or governments, for example, will sometimes do a better job.

- *Repugnant markets*: Marketing some goods and services—vital organs, or human beings—violates an ethical norm, or undermines the dignity of those involved.

Other institutions may be more effective

In Unit 6 we found that the allocation of work and other resources *within firms* is not accomplished by buying and selling, but instead by the authority of the owners and their managers and the directives they give to subordinates.

Firms exist because it is more profitable for the owners to organise production in that way. In this case, the extent of the market is determined by the firm's decision about which components of a product to produce and which ones to buy. The extent of the market is determined entirely by the owners' considerations of the more profitable way to do business.

There is another way that markets may not be the best way to get the job done. Markets work, in part, because prices provide incentives, and changes in prices will motivate people to change their behaviour in a direction that improves overall economic performance.

But this is sometimes not true.

It is common for parents to rush to pick up their children from daycare. Sometimes a few parents are late, making some teachers stay extra time. What would you do to deter parents from being late? In 2000, economists ran an experiment introducing fines in some daycare centres in Israel. The "price of lateness" went from zero to ten Israeli shekels (about $3 at the time). Surprisingly, after the fine was introduced, the frequency of late pickups doubled. The top line in Figure 10.9 illustrates this.

Why did putting a price on lateness backfire?

One possible explanation is that before the fine was introduced, most parents were on time because they felt that it was the right thing to do. In other words, they came on time because of a moral obligation to avoid inconveniencing the daycare staff. But the imposition of the fine signalled that the situation was really more like shopping. Lateness had a price and so could be purchased like a breakfast cereal.

The use of a market-like incentive—the price of lateness—had provided what psychologists call a new *frame* for the decision, making it one in which self-interest rather than concern for others was acceptable. When fines and prices have these unintended effects we say that incentives have *crowded out* social preferences. Even worse, you can also see from Figure 10.9 that when the fine was removed, parents continued to pick up their children late.

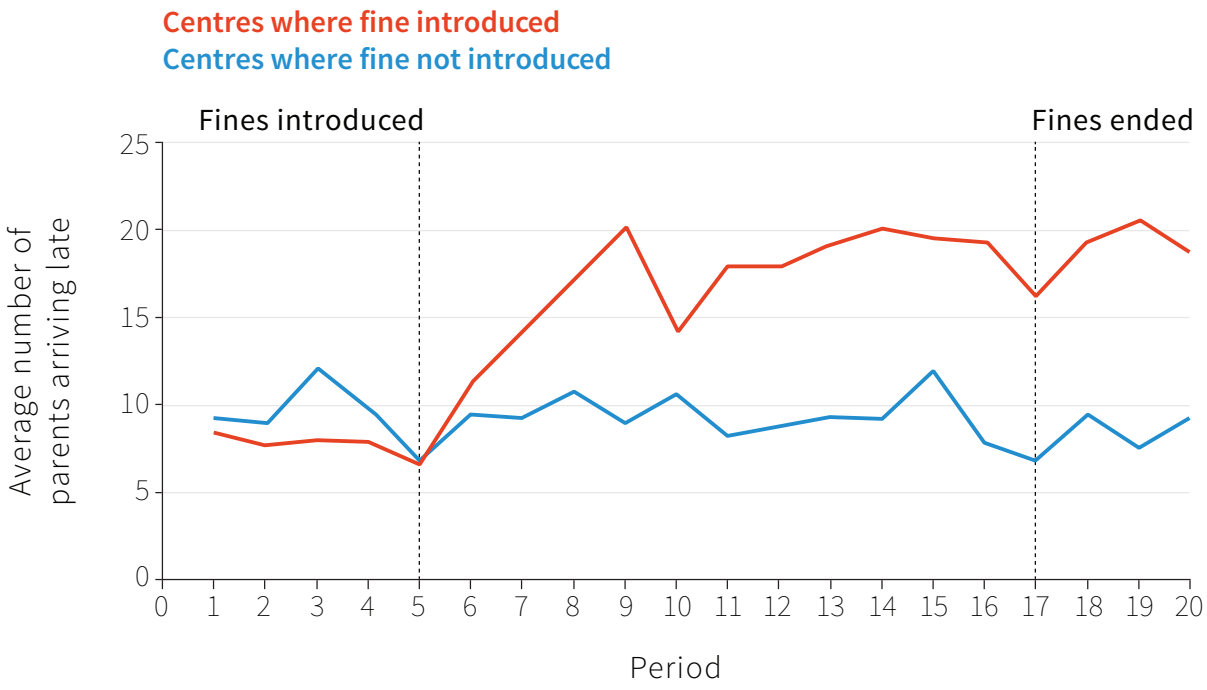


Figure 10.9 *Average number of late-coming parents, per week.*

Source: *Figure 1 from Gneezy, Uri, and Aldo Rustichini. 2000. 'A Fine Is a Price.' The Journal of Legal Studies 29 (January): 1–17.*

DISCUSS 10.9: CROWDING OUT

Imagine you are the mayor of a small town and wish to motivate your citizens to get involved in "City Beautiful Day". This would involve cleaning parks and roads.

How would you design the day to motivate citizens to take part?

Repugnant markets

In most countries primary education is provided free to all children (usually compulsorily) by the public sector. Philosophers have long argued that access to education is a right; it should be provided equally for all, and should not depend on willingness or ability to pay. These goods are sometimes called *merit goods*.

Buying and selling babies raises different issues. There are well-established institutions allowing parents to voluntarily give up a baby for adoption, but laws typically prevent parents from selling their children, and most people feel that adoption should not be a monetary transaction.

Most people think the same about the sale of human organs for transplant. But we can make the argument that it is wrong to prevent these transactions if both parties enter into them voluntarily. One reason we might object: the sale may not be truly voluntary because poverty might force people to enter into a transaction they might later regret. A second reason would be a belief that to put a price on a baby, or a body part, violates a principle of human dignity. It corrupts our attitudes towards others.

Alvin Roth, an economist who won a Nobel prize for his work, calls these repugnant markets.

The philosophers Michael Walzer and Michael Sandel have discussed the moral limits of markets (you can watch Sandel investigating the moral limits of his audience in this video). Some market transactions conflict with the way we value humanity; others with principles of democracy, such as allowing people to sell their votes. We have seen some of the advantages of allocating resources using markets and the price system; in that analysis we implicitly assumed that exchanging the good for money did not affect its intrinsic value to the buyer and seller.

But parents' attitudes to babies, and voters' appreciation of their democratic rights, might both be altered if they were bought and sold. When we consider whether it would be beneficial to introduce a new market, or monetary incentives, we should think about whether this might crowd out other social norms or ethical preferences.

DISCUSS 10.10: CAPITALISM AMONG CONSENTING ADULTS

Should all voluntary contractual exchanges be allowed among consenting adults?

What do you think about the following (hypothetical) exchanges? You may assume in each case that the people involved are sane, rational adults who have thought about the alternatives and consequences of what they are doing. In each case, decide whether you approve, and whether you think the transaction should be prohibited.

1. A complicated medical procedure has been discovered that cures a rare form of cancer in patients who would otherwise certainly die. Staff shortages make it impossible to treat all those who would benefit, and the hospital has established a policy of first come, first served. Ben, a wealthy patient who is at the bottom of the list, offers to pay Aisha, a poor person on the top of the list, $1m to exchange places. If Aisha dies (which is very likely), then her children will inherit the money. Aisha agrees.

2. Melissa is 18. She has been admitted to a good university but does not have any financial aid, and cannot get any. She signs a four-year contract to be a stripper on the internet and will begin work when she is 19. The company will pay her tuition fees.
3. Space Marketing Inc. announces plans to launch giant billboards made from mylar sheets into low orbit. Companies would pay more than $1m dollars to display advertisements. Logos, about the size of the moon, will be visible to millions of people on earth.
4. You are waiting in line to buy tickets for a movie that is almost sold out. Someone from the back of the line approaches the person in front of you and offers her $25 to let him in front of her.
5. A politically apathetic person, who never votes, agrees to vote in an election for the candidate who pays him the highest amount.
6. William and Elizabeth are a wealthy couple who give birth to a baby with a minor birth defect. They sell this baby to their (equally wealthy) neighbours and buy a child without any birth defects from a family who need the money.
7. A care home for elderly people advertises for nurses, saying "Jamaicans preferred". The director justifies it by saying: "In our experience, Jamaican nurses are the most efficient".
8. A well-informed and sane adult, with an adequate income, decides that he would like to sell himself to become the slave of another person. He finds a buyer willing to pay his asking price. The aspiring slave will give the price paid by the buyer to his children to further their education.

10.11 ANOTHER SOURCE OF MARKET FAILURES: $P > MC$

Even in the absence of environmental external effects, public goods, and the other sources of market failure studied so far in this unit, market failures may arise when firms set prices that exceed marginal cost. In this case the allocation is not Pareto efficient, as we saw in Unit 7. This may occur for two reasons:

- *Limited competition*: If all actors are price-takers in the equilibrium of a competitive market, firms will sell at prices equal to marginal costs ($P = MC$). The price is thus a true signal of the scarcity of the good, and the resulting allocation is Pareto efficient. But In Unit 7 we saw that firms that are facing limited competition—monopolists or those producing differentiated goods—set their

prices above marginal cost. The price at which the good is sold then sends the wrong message because the high price overstates the real scarcity of the good as indicated by its marginal cost. The resulting allocation is not Pareto efficient: too little is sold, so there is a deadweight loss.

- *Decreasing long-run average costs*: Firms may not be able to set a price equal to marginal cost and cover their average costs. This will occur when the firm's long-run average cost curve is downward-sloping. As more units are produced, the average cost per unit produced gets smaller. This will be the case if the production process is characterised by economies of scale or if the price of inputs declines as the firm purchases larger quantities. If the cost curve is downward-sloping then the marginal cost must be less than the average cost (as in Figure 7.5b, the cost curve for Apple-Cinnamon Cheerios), and the firm cannot set a price equal to the marginal cost that also covers average costs. Extreme cases of this problem are sometimes referred to as natural monopolies, for example power grids and transportation networks.

Economies of scale that create declining unit costs are a common market failure. Think about a film production company. The company spends heavily on hiring actors, camera technicians, a director, purchasing rights to the script and advertising the film. These are the firm's *fixed costs* (sometimes called *first copy costs*). The cost of making available additional copies of the film (the marginal cost) is typically low: the first copy is cheap to reproduce. This firm's marginal costs will be below its average costs (including the normal rate of profit). If it were to set a price equal to marginal cost it would go out of business. Notice that the problem here is *not* lack of competition—the film industry is highly competitive—but the fact that the marginal cost curve is always lower than the average cost curve, meaning that the long-run average cost curve is downward-sloping.

However, the two problems—limited competition and decreasing long-run average costs—are often closely related because competition among firms with downward-sloping average cost curves tends to be winner-takes-all. The first firm to exploit the cost advantages of large size eliminates other firms and, as a result, eliminates competition too.

But, in both these cases, too little is purchased: there are some potential buyers whose willingness to pay exceeds the marginal cost but falls short of the price set by the firm—so they won't buy the good. As a result, there is a deadweight loss. To see why, suppose hypothetically that the firm knew each of its potential buyers' maximum willingness to pay for the product, and could charge *separate prices* to each, just below the buyer's willingness to pay. Then the firm would definitely sell to any potential buyer whose willingness to pay exceeded the marginal cost, and as a result all potentially beneficial trades would take place.

If the firm could set individual prices for each buyer—called *price discrimination*—then the deadweight loss would disappear. Note that the firm would capture the entire surplus (there would be only producers' surplus, no consumers' surplus). You

might think this unfair, but the market allocation would be Pareto efficient because no further mutually beneficial payments, special prices or any other voluntary exchanges would be possible.

We can use this thought experiment, in which the hypothetical firm sets individual prices for each buyer, when we think about the problem of a monopoly. It demonstrates that there is a market failure, and helps us to understand why it occurs.

But in reality, firms are no more able to charge a specially tailored price to each potential buyer than the fisheries in Martinique are able to make payments to each plantation to reduce the use of Chlordecone. Neither the firms, nor the fisheries nor the courts could possibly have the necessary information to write the necessary contracts, much less to enforce them.

The decisions by firms facing limited competition or with high fixed costs affect others, as we can see below in the final line of our table (in the next section you can see the entire table that we have built, line by line, in this unit). When $P > MC$, the external effects of the firms' decisions produce market failure. As usual we show the possible remedies in the penultimate column although, in this case, some of them are clearly impractical.

DISCUSS 10.11: KNOWLEDGE PRODUCTION AND NATURAL MONOPOLY

Some firms in the knowledge production sector have very high fixed costs because of the patents, copyrights or trademarks which they own. Their long-run average cost curve is downward-sloping. Are such firms *natural monopolies*?

THE DECISION	HOW IT AFFECTS OTHERS	COST OR BENEFIT	MARKET FAILURE (MISALLOCATION OF RESOURCES)	POSSIBLE REMEDIES	TERMS APPLIED TO THIS TYPE OF MARKET FAILURE
The firm sets the price of the good so that $P > MC$	Some would - be buyers do not buy	Foregone benefit to seller and buyer	Some whose willingness to pay exceeds MC do not buy the good	Competition policy, price discrimination, public ownership of natural monopolies, subsidies and other policies to offset the firm's fixed costs that account for declining AC	Limited competition (downward - sloping firm demand curve), or high fixed costs, (decreasing long run average costs)

10.12 CONCLUSION

Anything that we care about can be called a good (or if we dislike it, a bad). Economics is about how goods (and bads) are allocated between people, and one means of allocation is trade in a market. Imagine that two economics students are discussing what they have learnt about markets in Units 6 to 10:

Marianna If everything we cared about was allocated by trade in a competitive market where everyone was a price-taker, the allocation of goods would be Pareto efficient. We need to make sure that there are property rights for all goods.

Caterina But that's not much use. We know that in most markets firms can set prices.

Marianna But still, if goods are traded in markets the prices give at least an indication of scarcity, and that helps to make sure they are allocated in the right way. And markets give people incentives to innovate and produce better or cheaper goods.

Caterina Problem is, there are lots of goods that can't be traded in markets at all. Think about the fishermen affected by the run-off of effluent from the banana plantation—they care about the quality of the water in the sea, and there's no market for that. And that means there are external effects. And what about R&D? You need a market for new knowledge, but that won't work because it's a public good.

Marianna That's where Coase comes in. Where there isn't a market to do the job you establish property rights so that, when the allocation of goods is inefficient, people can bargain and write contracts to sort out the problem.

Caterina I suppose that might work for some things. But it isn't always possible to write contracts that can be upheld in court—it's hard to imagine solving road congestion or innovation incentives that way. I suppose you'd argue that roads ought to be privately owned.

Marianna But people find other ways of solving social dilemmas—like irrigation systems or writing open source software. Perhaps markets aren't always the answer. Anyway, we've forgotten something important.

Caterina Fairness?

Marianna Yes. It's great if markets and property rights help us allocate goods efficiently, but they won't help us distribute them fairly. Maybe that's the biggest social dilemma of all.

At this point, the two students have reached some agreement. We know from the story of Walrasia that looking at the markets in the economy can help us understand why some people are rich and others are poor. If some people have valuable endowments and others don't, and we consider this to be unfair, markets alone will not address the problem. We will return to this problem in Unit 19.

Marianna imagines an "ideal" world in which everything we care about is allocated in markets (preferably competitive ones) or, where markets are not possible, we can still exchange goods for money by writing contracts enforceable in court. Economists often think about worlds like this: not because it is feasible or desirable, but because it helps us understand the problems of market failure associated with the allocation of individual goods.

For example, what Caterina says about the fishermen is true: one way of interpreting an external effect is to say that, *if there were a market* in which fishermen and plantation owners *could* trade rights to clean water, it would solve the problem. In the case of the Caribbean fisheries that probably isn't feasible, but we will see in Unit 18 that new markets have been established to address the problem of carbon emissions.

We also imagined a firm that could write separate contracts with each of its customers, setting a price exactly equal to the customer's willingness to pay, to show why limited competition creates a market failure. This case—because it is unrealistic—shows why a market failure is likely: fine-tuning in price discrimination is usually impossible.

CONCEPTS INTRODUCED IN UNIT 10

Before you move on, review these definitions:

- *Market failure*
- *External effect (externality)*
- *Incomplete contract*
- *Asymmetric information*
- *Verifiable information*
- *Pigouvian tax (or subsidy)*
- *Coasean bargaining*
- *Marginal social cost*
- *Public good*
- *Positional good*
- *Veblen effect*
- *Crowding out*
- *Repugnant markets*
- *Merit good*
- *Patent*

Market failures affect many of the activities observed in a capitalist economy and economists give them different names. Market failures have a common underlying structure, which we can capture by asking a series of questions:

QUESTION	ANSWER
Why do they happen?	People, guided only by market prices, do not take account of the full effect of their actions on others
Why is the full effect of their actions on others not taken into account?	There are external benefits and costs that are not compensated by payments
Why do market prices not do the job?	Market prices will not measure the entire social costs and benefits of goods and services
Why can't private bargaining and payments solve the problem?	This requires property rights and contracts, which are absent or unenforceable by courts
What prevents the necessary property rights and contracts from being written and upheld in a court of law?	Asymmetric or non - verifiable information

Figure 10.10 *Market failures and information problems.*

Figure 10.10 makes it clear that information asymmetries, and the way these limit the kinds of contracts that can be written, are the source of market failures. The same information problems can hamper a government seeking to use taxes, subsidies, or prohibitions to achieve the Pareto-efficient outcome. For example, the governments of Guadaloupe and Martinique eventually decided to ban the use of Chlordecone rather than to try to tax banana production or provide compensation to the fisheries.

Sometimes a combination of remedies is the best way to cope with these information asymmetries, for example, in providing car insurance. In many countries, third party insurance (covering damage to others) is compulsory to avoid the adverse selection problem that would occur if only the accident-prone drivers purchased insurance. To address the moral hazard problem of hidden actions, however, insurers sometimes require the installation of on-board monitoring devices so that prudent driving habits can be an enforceable part of the contract.

In the earlier sections of this unit we have built up a table of market failures, and the possible remedies for them. So, in Figure 10.11, we can collect all our examples of market failure in one table:

THE DECISION	HOW IT AFFECTS OTHERS	COST OR BENEFIT	MARKET FAILURE (MISALLOCATION OF RESOURCES)	POSSIBLE REMEDIES	TERMS APPLIED TO THIS TYPE OF MARKET FAILURE
A firm uses a pesticide that runs off into waterways	Downstream damage	Private benefit, external cost	Overuse of pesticide and overproduction of crop in which it is used	Taxes, quotas, bans, bargaining, common ownership of all affected assets	Negative external effect, environmental spillovers *(Section 10.1)*
You take an international flight	Increase in global carbon emissions	Private benefit, external cost	Overuse of air travel	Taxes, quotas	Public bad, negative external effect *(Section 10.5)*
You travel to work by car	Congestion for other road users	Private cost, external cost	Overuse of cars	Tolls, quotas, subsidised public transport	Common pool resource, negative external effect *(Section 10.5)*
A firm invests in R&D	Other firms can exploit the innovation	Private cost, external benefit	Too little R&D	Publicly funded research, subsidies for R&D, patents	Public good, positive external effect *(Section 10.5)*
A firm trains a worker in a skill useful in other firms	Another firm benefits if the worker switches jobs	Private cost, external benefit	Too little training	Subsidies for firms that train	Public good, positive external effect *(Section 10.4)*
An employee on a fixed wage decides how hard to work	Hard work increases her employer's profits	Private cost, external benefit	More effort and higher wages would be better for both worker and employer	More effective monitoring to make the contract more complete, reduced conflict of interest between employer and worker	Incomplete labour contract (does not cover effort), hidden action, moral hazard *(Section 10.8)*
An unemployed worker offers to work as hard as an employed worker at a lower wage	Would confer a benefit on the employer	External benefit	Despite this being a mutually beneficial agreement, it could not be enforced so the employer refuses, involuntary unemployment	More effective monitoring to make the contract more complete, reduced conflict of interest between employer and worker	Incomplete labour contract (does not cover effort), hidden action, moral hazard *(Section 10.8)*
A person who knows he has a serious health problem decides to purchase insurance	Results in insurance provider making losses	Private benefit, external cost	Those whose higher risk exposure is known only to them (not the insurance company) will be more likely to buy	Mandatory purchase of health insurance, public provision, mandatory health information sharing	Missing markets (hidden attribute, adverse selection) *(Section 10.8)*
A person who has purchased car insurance takes more risks	More prudent driving would contribute to insurance company profits	Private benefit to the insured, external cost for the insurance company	Insurance is more expensive than it would be were there no hidden actions, too little insurance purchased	Installing monitoring devices that generate verifiable information on driving habits	Missing markets (hidden action, moral hazard) *(Section 10.8)*
The firm sets the price of the good so that $P > MC$	Some would - be buyers do not buy	Foregone benefit to seller and buyer	Some whose willingness to pay exceeds *MC* do not buy the good	Competition policy, price discrimination, public ownership of natural monopolies, subsidies and other policies to offset the firm's fixed costs that account for declining *AC*	Limited competition (downward - sloping firm demand curve), or high fixed costs, (decreasing long run average costs) *(Section 10.11)*

Figure 10.11 *Some cases of market failure.*

DISCUSS 10.12: MARKET FAILURE

Construct a table like the one in Figure 10.11 to analyse the possible market failures associated with the decisions below. In each case can you identify which markets or contracts are missing or incomplete?

1. You buy an item of luxury designer clothing
2. You inoculate your child with a costly vaccination against an infectious disease
3. You use money that you borrow from the bank to invest in a highly risky project
4. A fishing fleet moves from the overfished coastal waters of its own country to international waters
5. A city airport increases its number of passenger flights by allowing night-time departures
6. You contribute to a Wikipedia page
7. A government invests in research in nuclear fusion

Key points in Unit 10

Pareto inefficiency

Pareto-inefficient market outcomes result if:

- Competition is limited
- Long-run average costs decline with output so that marginal cost is always less than average cost
- Some aspect of the exchange is not covered by an enforceable property right or contract

Asymmetric and non-verifiable information

Asymmetric and non-verifiable information relevant to the exchange make it impossible to establish enforceable property rights and for contracts to cover all aspects of an exchange. Examples include:

- Public goods such as knowledge
- Public bads such as Veblen goods and environmental spillovers

Unexploited opportunities for exchange

If a market outcome leaves unexploited opportunities for mutually beneficial exchanges, the resulting Pareto inefficiency means that the market has failed.

Coasian bargaining and Pivouvian taxes may improve outcomes

Both Coasian bargaining and Pigouvian taxes and subsidies can improve on market outcomes in these cases; but both are limited by the same problems of asymmetric and non-verifiable information that is the reason for the market failure.

The trade-off between incentivising innovation and diffusion

Policies to encourage development and use of new knowledge (new technologies, new products) face a trade-off between:

- Providing incentives to create the innovation (for example, patents)
- The rapid and widespread diffusion of new ideas (obstructed by the monopoly rights of patent-holders)

Markets have limits

Repugnance and other moral objections to exchanging some goods for money, and the crowding-out effects of monetary incentives, provide reasons why not all goods and services are allocated on markets.

10.13 EINSTEIN

Endowments and inequality in the equilibrium of price-taking markets

To see why the process of exchange on a market in Walrasia does not introduce any additional inequalities not already present in the disparities in the islanders' endowments, think about a miner's family from Coal Island and a farmer's family from Wheat Island.

The price of coal in the market is $70 per tonne and the price of wheat is $280 per tonne.

The miner brings a tonne of coal to the market, which is half of his production since the last market, and sells it for $70. He uses the proceeds to buy a quarter of a tonne of wheat. He is better off because of the exchange. We know he is better off because he chose to sell the coal and buy the wheat—he could have chosen to keep his coal. Now he has enough coal to heat his home and plenty of wheat to make bread.

A farmer from Wheat Island goes to the market. She produced a tonne of wheat since her previous visit. She sells a quarter of a tonne of wheat and buys a tonne of coal. She is also better off, for the same reason.

We now compare the situation before and after trading in the market. The wheat farmer started out with a tonne of wheat worth $280 and ended up with three quarters of a ton of wheat worth $210 and half a ton of coal worth $70. So she ended up with $280 worth of goods, exactly the value of the grain she had started out with.

The miner started out with two tonnes of coal worth $70 each or a total of $140. He ended up with a tonne of coal (worth $70) and a quarter of a tonne of grain (worth $70) for a total of $140. This is exactly the value of the coal that he started out with.

In this case access to the market:

- Increased utility for each participant
- Did not change the monetary value of the things that the two traders possessed: they ended up with $140 for the miner and $280 for the farmer, exactly what they had to begin with.

So the market did not create inequality. If the exchanges that took place were part of the equilibrium of a price-taking market, the only effect is to allow the things owned by individuals' specialised production (coal, wheat) to be transformed into a mixed set of goods (some coal, some wheat for each), which they prefer.

10.14 READ MORE

Bibliography

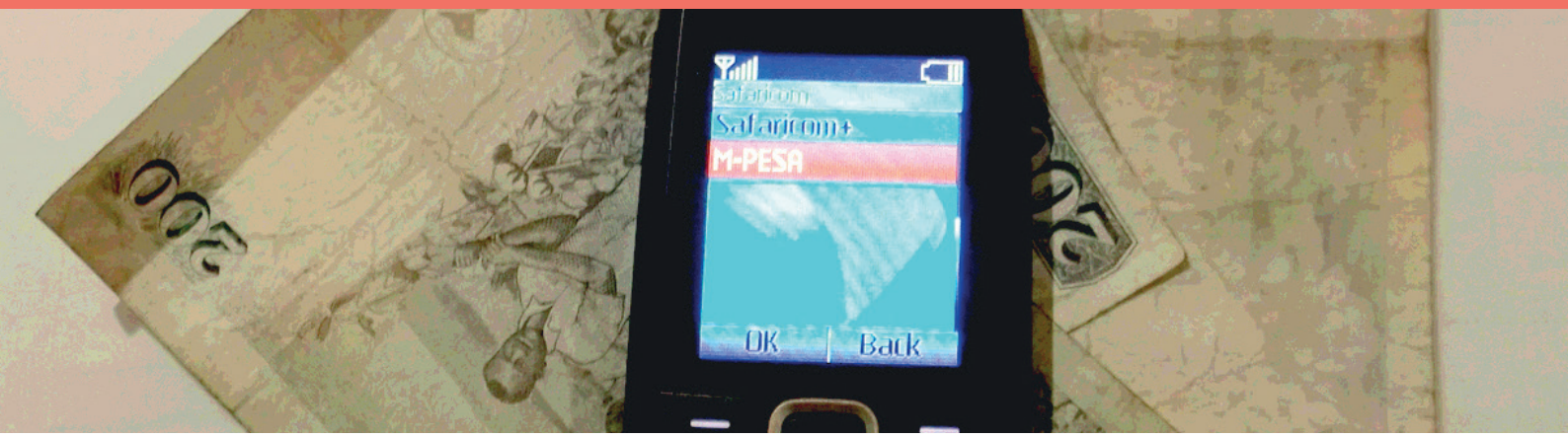
1. Acemoglu, Daron, Simon Johnson, and James A. Robinson. 2005. 'Institutions as a Fundamental Cause of Long-Run Growth.' In *Handbook of Economic Growth*, Volume 1A, edited by Philippe Aghion and Steven N. Durlauf. Amsterdam: North Holland.
2. Acemoglu, Daron, and James A. Robinson. 2012. *Why Nations Fail: The Origins of Power, Prosperity and Poverty*. 1st ed. New York, NY: Crown Publishers.
3. Akerlof, George A., and Robert J. Shiller. 2015. Phishing for Phools: The Economics of Manipulation and Deception. Princeton, NJ: Princeton University Press.
4. Alesina, Alberto, and Eliana La Ferrara. 2000. 'Participation in Heterogeneous Communities.' *Quarterly Journal of Economics* 115 (3): 847–904.
5. Bowles, Samuel. 2016. *The Moral Economy: Why Good Incentives Are No Substitute for Good Citizens*. New Haven, CT: Yale University Press.
6. Clark, Andrew E., Paul Frijters, and Michael A. Shields. 2008. 'Relative Income, Happiness, and Utility: An Explanation for the Easterlin Paradox and Other Puzzles.' *Journal of Economic Literature* 46 (1): 95–144.
7. Edison Innovation Foundation. 2015. 'All about Tom.' *Thomasedison.org*.
8. Fafchamps, Marcel, and Bart Minten. 1999. 'Relationships and Traders in Madagascar.' *Journal of Development Studies* 35 (6): 1–35.
9. Frank, Robert H. 2011. 'The Progressive Consumption Tax: A Win-Win Solution for Reducing American Economic Inequality.' *Slate*. December 7.
10. Glaeser, Edward L. 2009. 'The Lorax Was Wrong: Skyscrapers Are Green.' *Economix*, March 10.
11. Gneezy, Uri, and Aldo Rustichini. 2000. 'A Fine Is a Price.' *The Journal of Legal Studies* 29 (January): 1–17.
12. Keynes, John Maynard. 1936. *The General Theory of Employment, Interest and Money*. London: Palgrave Macmillan.
13. Mazzucato, Mariana. 2011. *The Entrepreneurial State*. London: Demos.
14. Mill, John Stuart. (1848) 1994. *Principles of Political Economy*. New York: Oxford University Press.
15. North, Douglass C. 1990. *Institutions, Institutional Change and Economic Performance*. Cambridge: Cambridge University Press.
16. Oh, Seung-Yun, Yongjin Park, and Samuel Bowles. 2012. 'Veblen Effects, Political Representation, and the Reduction in Working Time over the 20th Century.' *Journal of Economic Behavior & Organization* 83 (2): 218–42.
17. Pigou, Arthur. 1912. *Wealth and Welfare*. London: Macmillan & Co.

18. Pigou, Arthur. 1920. *The Economics of Welfare*. London: Macmillan & Co.
19. Pigou, Arthur. 1933. *Theory of Unemployment*. London: Macmillan & Co.
20. Roth, Alvin E. 2007. 'Repugnance as a Constraint on Markets.' *Journal of Economic Perspectives* 21 (3): 37–58.
21. Sandel, Michael. 2009. *Justice*. London: Penguin.
22. Schor, Juliet B. 1991. *The Overworked American: The Unexpected Decline of Leisure*. New York, NY: Basic Books.
23. Seabright, Paul. 2010. *The Company of Strangers: A Natural History of Economic Life* (Revised Edition). Princeton, NJ: Princeton University Press.
24. Smith, Adam. (1776) 2003. 'That the Division of Labour Is Limited by the Extent of the Market.' In *An Inquiry into the Nature and Causes of the Wealth of Nations*, by Adam Smith. New York, NY: Random House Publishing Group.
25. *The Economist*. 2015. 'A Question of Utility.' August 8.
26. *The Economist*. 2015. 'Time to Fix Patents.' August 8.
27. Walzer, Michael. 1983. *Spheres of Justice: A Defense of Pluralism and Equality*. New York, NY: Basic Books.

Beta December 2015 version

11

CREDIT, BANKS AND MONEY



HOW CREDIT, BANKS AND MONEY EXPAND OPPORTUNITIES FOR MUTUAL GAIN, AND WHAT LIMITS THEIR CAPACITY TO ACCOMPLISH THIS

- People can rearrange the timing of their spending by borrowing, lending, investing and saving
- While mutual gains motivate credit market transactions, there is a conflict of interest between borrowers and lenders over the rate of interest, the prudent use of loaned funds and their repayment
- People with limited wealth are sometimes unable to secure loans, or can do so only at high rates of interest, or if the project they wish to finance is small or exceptionally productive
- Money is a medium of exchange consisting of bank notes, cheques, credit—or anything else that can be used to purchase things—that is accepted as payment because others can use it for the same purpose
- Banks are profit-maximising firms that create money in the process of supplying credit
- A nation's central bank is normally part of government, and produces money by issuing legal tender and lending to banks at its chosen policy rate of interest
- The interest rate charged by banks to borrowers (firms and households) is determined in large measure by the policy interest rate chosen by the central bank

Institute for New Economic Thinking
UCL SciencesPo
Azim Premji University
FRIENDS PROVIDENT Foundation

The market town of Chambar in south-eastern Pakistan serves as the financial centre for 2,400 farmers in surrounding villages. At the beginning of the *kharif* planting season in April, when they sow cotton and other cash crops, they will buy fertiliser and other inputs. Months have passed since they sold the last harvest and so the only way they can do this is to borrow, promising to repay at the next harvest. Others borrow to pay for medicines or doctors. But few of them have ever walked through the shiny glass and steel doors of the J.S. Bank on Tando Allahyar Road. Instead they visit one of approximately 60 moneylenders.

If they are a seeking a first-time loan they will be questioned intently by the moneylender, asked for references from other farmers known to the lender, and in most cases given a small trial loan as a test of creditworthiness. The lender will probably visit to investigate the condition of the farmer's land, animals and equipment.

The lenders are right to be wary. If the farmer's crop fails due to the farmer's lack of attention, the lender loses money. Unlike many financial institutions, lenders do not usually require that the farmer set aside some property (called *collateral*: for example, some gold jewellery) that would become the lender's property if the farmer were unable to repay the loan.

COLLATERAL

Collateral is a borrower's asset the ownership of which will be transferred to the lender if the borrower fails to repay the loan as agreed.

If the would-be first time borrower looks like a good risk, he will be offered a loan. In Chambar, this is at an average interest rate of 78% per annum. Loans in Chambar are typically for four months (the growing period of the crop prior to harvest), so 100 rupees borrowed before planting will be paid back as 126 rupees. Knowing that more than half the loan applications are refused, he would consider himself fortunate.

And indeed he would be, at least compared to some people 12,000km away in New York who take out short-term loans to be repaid when their next pay cheque comes in. These payday loans bear interest rates between 350% and 650% per annum. The legal maximum interest rate in New York is 25%. In 2014 the "payday syndicate" offering these loans was charged with criminal usury in the first degree.

Given the interest rate, is lending in Chambar likely to be exceptionally profitable? The evidence suggests it is not. Some of the funds lent to farmers are borrowed from commercial banks like the J.B. Bank, at interest rates averaging 32% per annum, representing a cost to the moneylenders. And the costs of the extensive screening and collection of the debts further reduces the profits made by the lenders.

Partly as a result of the careful choices made by the moneylenders, default is rare: fewer than one in 30 fails to repay. By contrast, default rates on loans made by commercial banks are much higher: one in three. The moneylenders' success in avoiding default is based on their accurate assessment of the likely trustworthiness of their clients.

Money and trust are more closely related than you might think.

On 4 May 1970, a notice appeared in the *Irish Independent* newspaper in the Republic of Ireland, titled "Closure of Banks". It read:

"As a result of industrial action by the Irish Bank Officials' Association… it is with regret that these banks must announce the closure of all their offices in the Republic of Ireland… from 1 May, until further notice."

Banks in Ireland did not open again until 18 November, six and a half months later.

Did Ireland fall off a financial cliff? To everyone's surprise, far from collapse, the Irish economy continued to grow much as before. A two-word answer has been given to explain how this was possible: Irish pubs. Andrew Graham, an economist, visited Ireland during the bank strike and was fascinated by what he saw:

"Because everyone in the village used the pub, and the pub owner knew them, they agreed to accept deferred payments in the form of cheques that would not be cleared by a bank in the near future. Soon they swapped one person's deferred payment with another thus becoming the financial intermediary. But there were some bad calls and some pubs took a hit as a result. My second experience is that I made a payment with a cheque drawn on an English bank (£1 equalled 1 Irish punt at the time) and, out of curiosity, on my return to England, I rang the bank (in those days you could speak to someone you knew in a bank) and they told me my cheque had duly been paid in but that on the back were several signatures. In other words, it had been passed on from one person to another exactly as if it were money".

The Irish bank closures are a vivid illustration of the definition of money: it is anything accepted in payment. Therefore, the amount of money in the economy is the amount that can be spent, so it includes, for example, credit in the form of overdrafts. At the time, notes and coins made up about one-third of the money in the economy, with the remaining two-thirds in bank accounts and credit. Credit cards were not widely in use. The majority of transactions used cheques, but paying by cheque requires banks to ensure that people have the funds to back up their paper payments.

Writing a cheque in exchange for goods and services transfers purchasing power from buyer to seller. In a functioning banking system the cheque is cashed at the end of the day, and the bank credits the current account of the shop. If the writer of the cheque does not have enough money to cover the amount, the bank bounces the cheque, and the shop owner knows immediately that he has to collect in some other way. People generally avoid writing bad cheques as a result. Another example: if you

get a loan to buy a car, the bank credits your current account and you then write a cheque, use a credit card or initiate a bank transfer to buy the car. This is money in a modern economy.

So what happens when the banks close their doors and everyone knows that cheques will not bounce, even if the cheque writer has no money? Will anyone accept your cheques? Why not just write a cheque to buy the car when there is not enough money in your current account or in your approved overdraft? If you start thinking like this, you would not trust someone offering you a cheque in exchange for goods or services. You would insist on being paid in cash. But there is not enough cash in circulation to finance all of the transactions that people need to make. Everyone would have to cut back, and the economy would suffer.

How did Ireland avoid this fate? As we have seen, it happened at the pub. Cheques—even those written using chequebooks printed during the strike by local printers—were accepted in payment as money, because of the trust generated by the pub owners. Publicans (owners of the pubs) spend hours talking and listening to their patrons. They were prepared to accept cheques, which could not be cleared in the banking system, as payment from those judged to be trustworthy. During the six-month period that the banks were closed, about £5bn of cheques were written by individuals and businesses, and not processed by banks. It helped that Ireland had one pub for every 190 adults at the time. With the assistance of pubs and shops, cheques could circulate as money. The citizens of Ireland created the amount of new money needed to keep the economy growing during the bank closure.

Irish publicans and the moneylenders in the market town of Chambar would perhaps not recognise, among the many things they had in common, that they were creating money; and they would not know that in doing so they were providing a service essential to the functioning of their respective economies.

Not everyone passes the trustworthiness tests sets by pub owners and moneylenders, of course. And, in Chambar and New York, some of those who do pay much higher interest rates than others.

11.1 MONEY AND WEALTH

Borrowing and lending money, and the trust that makes this possible, are all about the passage of time. I trust you today, but I find out later if my trust was correct. The moneylender offers funds to the farmer to purchase fertiliser now, and he will pay

back after the crop matures, as long as it was not destroyed by a drought. The payday borrower will get her paycheque at the end of the month but needs to buy food now. She wants to bring some of her future buying power to the present.

The passage of time is an essential part of concepts such as money, income, wealth, consumption, savings and investment.

Money

Money is a medium of exchange including bank notes, cheques, bank deposits, or whatever else one can be used as a means of payment, which is accepted because others can use it for the same purpose. The "because" is important: it is what distinguishes exchange facilitated by money from barter exchange. In a barter economy I might exchange my apples for your oranges because I want some oranges, not because I intend to use the oranges to pay my rent. Money makes more exchanges possible because it's not hard to find someone who will be happy to have your money (in exchange for something) while unloading a large quantity of apples could be a problem. This is why barter plays a limited role in virtually all economies.

For money to do its work, it must be the case that almost everyone believes that if they accept money from you, in return for handing over their good or service, then they will then be able to use the money to buy something in turn. In other words they must trust that others will accept your money as payment. Governments and banks usually provide this trust. But the Irish bank closure shows that, when there is sufficient trust among households and businesses, money can function in the absence of banks. The publicans and shops accepted a cheque as payment, even though they knew it could not be cleared by a bank in the foreseeable future. As the dispute went on, the cheque presented to the pub or shop relied on a lengthening chain of uncleared cheques received by the person or business presenting the cheque. Some cheques circulated many times, endorsed on the back by the pub or shop owner, just like a bank note.

This establishes the fundamental characteristic of money as a medium of exchange. Money allows purchasing power to be transferred among people so that they can exchange goods and services, even when payment takes place at a later date (through the clearing of a cheque or settlement of credit card or trade credit balances, for example). Therefore money requires trust to function.

Wealth

A simple way to think about *wealth* is that, having paid off your debts and collected any money owed to you, it is the largest amount that you could consume without borrowing—for example if you sold your house, car and everything you owned.

The term wealth is also sometimes used in a broader sense to include immaterial aspects such as your health, skills and ability to earn an income (your human capital). But we will use the narrower definition in this unit.

Income

Income is the amount of profits, interest, rent, labour *earnings* and other payments, including those from the government. Income is a *flow* as it is measured over some period of time (annual income for example, or hourly wages). Wealth is a *stock*, meaning that it has no time dimension: at any moment of time it is just there.

To remember the difference between wealth and income, think of filling a bathtub, as in Figure 11.1. Wealth is the amount (stock) of water in the tub; income is the flow of water into the tub. The inflow is measured by litres (or gallons) per minute; the stock of water is measured by litres (or gallons) at a particular moment in time.

As we have seen, wealth often takes physical forms such as a house, or car, or office, or factory. The value of this wealth declines, either due to use or simply the passage of time.

This reduction in the value of a stock of wealth over time is termed *depreciation*. So for the bathtub, depreciation would be the amount of evaporation of the water. Like income, it is a flow (you could measure it in litres per year), but a negative one. So when we take account of depreciation we have to distinguish between *net income* and *gross income*: gross income is the flow into the bathtub, while net income is this less depreciation.

WEALTH

The term *wealth* includes the home that you own, your car, any land, buildings, machinery or other capital goods that you may own, and any financial assets such as shares or bonds. We subtract any debts that you have from this total (so that the wealth you have in your home is its market value, minus how much you owe to the bank). We add debts that others owe to you.

INCOME

The amount of profit, interest, rent, labour earnings and other payments (including transfers from the government) received, net of taxes paid, measured over a period of time such as a year. Your income is the maximum amount that you could consume and leave your wealth unchanged. It is also referred to as disposable income, to distinguish it from pre-tax income that is not all available to be spent.

Expenditure

The tub also has an outflow pipe or drain. The flow through the drain is called *consumption* expenditure, and it reduces wealth. In the bathtub, net income is equal to the maximal flow out the drain (consumption) such that, given the inflow from gross income and the evaporation of depreciation, the amount of water in the tub (wealth) is unchanged.

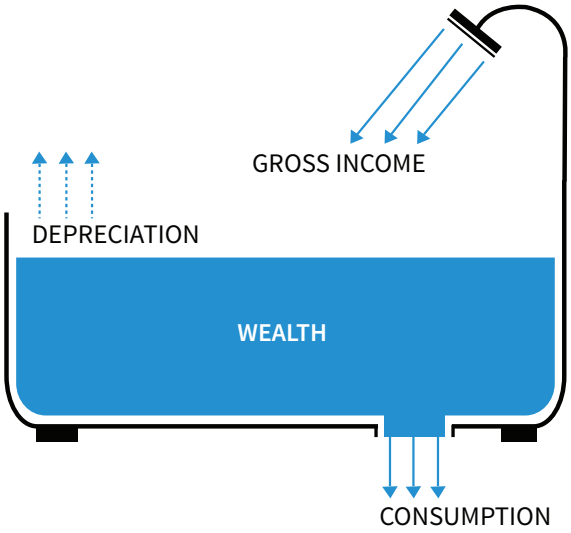


Figure 11.1 *Wealth, income, depreciation, and consumption: The bathtub analogy.*

An individual (or household) saves when consumption is less than net income, so her wealth increases. One form saving can take is the purchase of a financial asset such as shares (or stocks) in a company or a government bond. Although this is sometimes referred to as "investment" in everyday language, in economics, *investment* is expenditure on capital goods, that is on goods such as machinery or buildings or research that increases productive capacity. The purchase of a newly constructed house is also investment.

EARNINGS

Wages, salaries, and other income from labour.

DEPRECIATION AND NET INCOME

- *Depreciation* is the loss in value of a form of wealth that occurs either through use (wear and tear) or the passage of time (obsolescence).
- *Net income* is gross income minus depreciation.

The distinction between investment and purchasing shares or bonds is illustrated by a sole-proprietor business. At the end of the year, the owner decides what to do with her net income. Out of the net income, she decides on her consumption expenditure for the year ahead and saves the remainder. With her savings, she could buy shares or bonds or a bank deposit. On the other hand, she could instead spend on new assets to expand the business. This expenditure is investment.

More generally, *business investment is expenditure on productive equipment*.

11.2 BORROWING: BRINGING CONSUMPTION FORWARD IN TIME

To understand borrowing and lending we will use feasible set and indifference curve analysis. In Unit 3 and Unit 5 you studied how Alexei and Angela, faced with the trade-offs dictated by a feasible set makes choices between objectives such as free time on one axis, and grades or bushels of grain on the other. They made choices from the feasible set, based on preferences described by indifference curves representing how much they valued one objective relative to the other.

Here you will see that the same feasible set and indifference curve analysis applies to choosing between having something now, and having something later. Earlier we saw that giving up free time is a way of getting more goods, or grades, or grain. Now we see that giving up some good to be enjoyed now will sometimes allow us to have more goods later. The *opportunity cost* of having more goods now is having fewer goods later.

Borrowing and lending allow us to rearrange in time our capacity to buy goods and services. Borrowing allows us to buy more now, but constrains us to buy less later. To see how this works, think about Julia, who needs to consume now, but has no money today. She knows that in the next period (later), as a result of her pay cheque or harvest, she will have $100. Julia's situation is shown in Figure 11.2. Each point in the figure shows a combination of Julia's capacity to consume things, now and later. We assume that she spends what she has, so each point in the figure gives her consumption now (measured on the horizontal axis) and later (measured on the vertical axis).

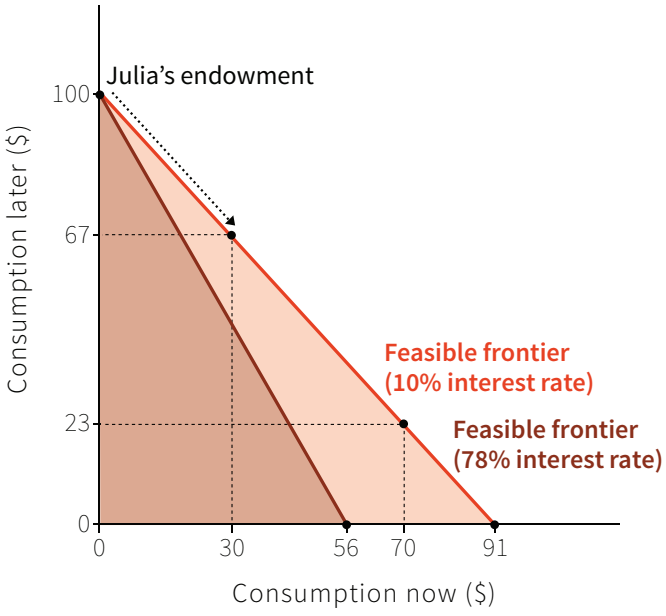


Figure 11.2 *Borrowing, the interest rate and the feasible set.*

Initially Julia is at the point labelled "Julia's endowment" in Figure 11.2. To consume now, Julia is considering taking out a payday loan (or she could be a farmer borrowing to finance her consumption before she can harvest and sell her crop).

Julia could, for example, borrow $91 now and promise to pay the lender the $100 that she will have later. Her total repayment of $100 would include the principal (how much she borrowed) plus the interest charge at the rate r, or:

$$repayment = principal + interest = 91 + 91r = 91(1 + r) = \$100$$

And if "later" means in one year from now, then the annual interest rate, r, is:

$$interest\ rate = \frac{repayment}{principal} - 1 = \frac{100}{91} - 1 = 0.10 = 10\%$$

You can think of the *interest rate* as *the price of bringing some buying power forward in time*.

At the same interest rate (10%), she could also borrow $70 to spend now, and repay $77 at the end of the year, that is:

$$repayment = 70 + 70r = 70(1 + r) = \$77$$

In that case she would have $23 to spend next year. Another possible combination is to borrow and spend just $30 now, which would leave Julia with $67 to spend next year, after repaying her loan.

All of her possible combinations of consumption now and consumption later ($91, $0; $70, $23; $30, $67 and so on) generate the feasible frontier shown in Figure 11.2, the boundary of the feasible set when the interest rate is 10%.

The fact that Julia can borrow means that she does not have to consume only in the later period. She can borrow now and choose any combination on her feasible frontier. But the more she consumes now, the less she can consume later. With an interest rate of $r = 10\%$, the opportunity cost of spending one dollar now is that Julia will have to spend $1.10 = 1 + r$ dollars less later.

One plus the rate of interest is the *marginal rate of transformation of goods from the future to the present* (to have one unit of the good now you have to give up $1 + r$ goods in the future.) This is the same concept as the marginal rate of transformation of free time into goods, grain or grades that you encountered in Units 3 and 5.

But suppose that, instead of 10%, the interest rate is now 78%, the average rate paid by the farmers in Chambar. At this interest rate Julia can only borrow a maximum of $56 now—because at 78% the interest on $56 comes to $44, using up all her $100 of future income. Her feasible frontier therefore shifts inward and the feasible set becomes smaller. Because the price of bringing buying power forward in time has

gone up, the capacity to consume in the present has fallen, just as your capacity to consume grain would fall if the price of grain went up (assuming you are not a producer of grain).

Of course the lender will benefit from a higher interest rate, as long as the loan is repaid, so there is a conflict of interest between the borrower and the lender.

11.3 IMPATIENCE AND THE DIMINISHING MARGINAL RETURNS TO CONSUMPTION

Given the opportunities for bringing forward consumption shown by the feasible set, what will Julia choose to do? How much she will bring forward will depend on how impatient she is. She could be impatient for two reasons:

- She prefers to smooth out her consumption instead of consuming everything later and nothing now.
- She may be an impatient type of person.

Smoothing

She would like to smooth her consumption because how much she enjoys of something she may consume will be greater when she has not already consumed a lot of it. Think about food: the first few bites of a dish are likely to be much more pleasurable than bites from your third serving. This is a fundamental psychological reality sometimes termed the *law of satiation of wants*.

The value to the individual of an additional unit of consumption declines the more that is consumed. This is called *diminishing marginal returns to consumption*. You have already encountered something similar in Unit 3, in which Alexei experienced diminishing marginal returns to free time: holding his grade constant, the more free time he had, the less each additional unit was worth to him, relative to how important the grade would be.

DIMINISHING MARGINAL RETURNS TO CONSUMPTION

The value to the individual of an additional unit of consumption declines, the more consumption the individual has. This is termed *diminishing marginal returns to consumption* (sometimes termed diminishing marginal utility).

Click on Figure 11.3a to see how diminishing marginal returns to consumption leads to the desire for consumption smoothing, and how this is represented by indifference curves.

As shown in Figure 11.3a Julia can choose her consumption now and later at any point on the dashed line. To find out what she chose, we added her indifference curves.

In Figure 11.3a we show two of a family of indifference curves that indicate Julia's preferences. She considers herself equally well off with any of the combinations of goods now and later represented by the points on one of the indifference curves. The shape of the indifference curves, bowed toward the origin, is a consequence of diminishing marginal returns to consumption. As with the other indifference curves in Units 3 and 5, the slope of the indifference curve is a *marginal rate of substitution (MRS)*, in this case between consumption now and consumption later. Note the slope of the curve is negative.

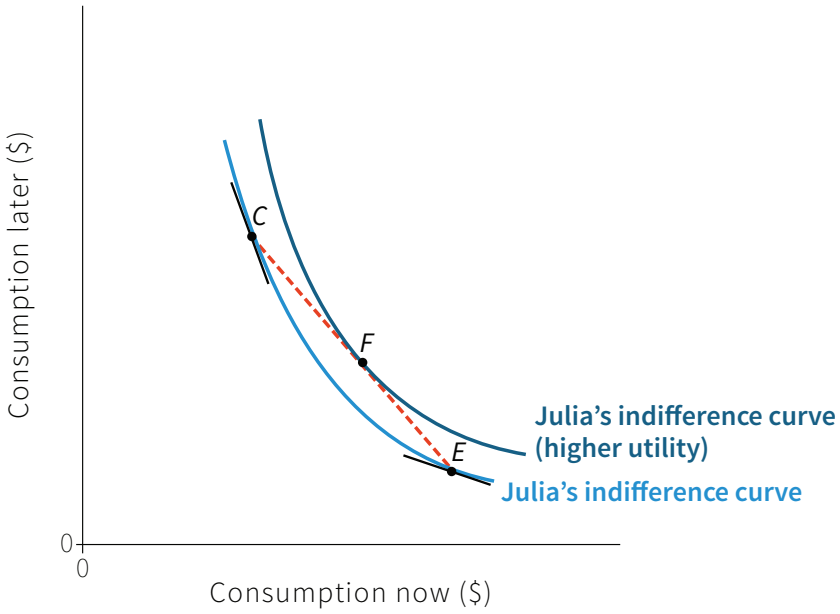


Figure 11.3a *Consumption smoothing: Diminishing marginal returns to consumption.*

We can see that she is equally happy with points C and E, which are on the same indifference curve—the first offering low consumption now and high later and the second offering the opposite.

- *The MRS at C is high* (look at the steep slope of her indifference curve): In the situation depicted by point C, Julia has little consumption now and a lot later, so diminishing marginal returns mean that she would like to move some consumption to the present.
- *The MRS at E is low*: She has a lot of consumption now and less later, so diminishing marginal returns mean that she would like to move some consumption to the future.

From this we know two things:

- She would not like to have a pattern of consumption that is either below *E* on the dashed line (even more now, even less later) or above *C*, the opposite.

- Julia would prefer some pattern of consumption between the "feast now, famine later" point E and the opposite at C.

Thus, diminishing marginal returns to consumption—now, and in the future—mean that Julia would like to smooth her consumption, that is, to pick some combination between C and E. But which point along the line CE will Julia choose?

If Julia is limited to combinations of C and E (that is, to points on the dashed line) the highest indifference curve available to Julia is the one with point F.

Pure impatience, or how impatient you are *as a person*

If Julia knows she can have two meals tomorrow but she has none today, then we have seen that diminishing marginal returns to consumption could explain why she might opt for the additional meal now—she prefers one meal today and one tomorrow. Note that Julia would opt for the meal now not because she is an impatient person but because she does not expect to be hungry in the future. She prefers to smooth her consumption of food.

But there is a different reason for preferring the good now, called *pure impatience*. To see whether someone is impatient *as a person*, we ask whether she values a good now more highly than later, even though she will have the same amounts in both periods. There are two reasons for pure impatience:

- *Myopia* (short-sightedness): People experience the present satisfaction of hunger or some other desire more strongly than they imagine the same satisfaction at a future date.
- *Prudence*: People know that they may not be around in the future, and so choosing present consumption may be a good idea.

"IMPATIENCE"

Any preference for present consumption over future consumption. This preference may be derived from:

- *Pure impatience*
- *Diminishing marginal returns to consumption*

To see what pure impatience means we compare two points on the same indifference curve. At point A she has \$50 now and \$50 later. We ask how much extra consumption she would need to have later in order to compensate her for losing \$1 now. Point B on the same indifference curve gives us the answer: if she had only \$49 now, she would need \$51.5 later in order to stay on the same indifference curve and be equally happy. So she needed \$1.5 later to compensate for losing \$1 now. Rather than preferring to perfectly smooth her consumption, she places more value on consumption today than in the future. Julia is an impatient person, meaning that she has pure impatience.

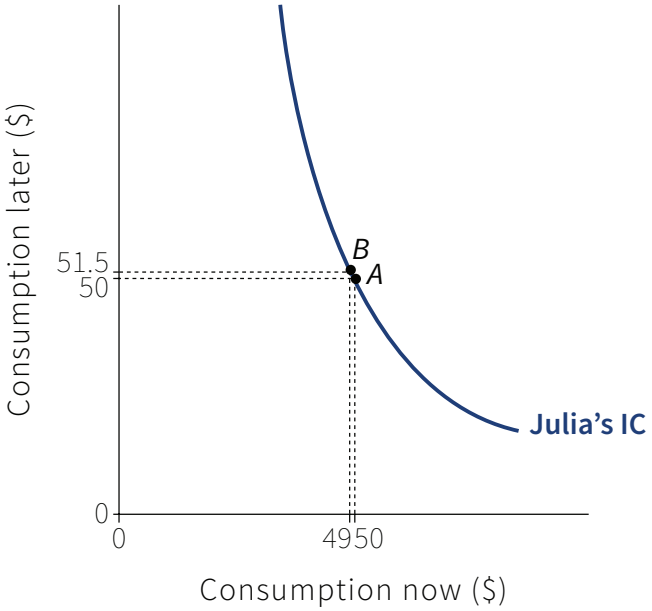


Figure 11.3b *Impatience and circumstance.*

The slope of the indifference curve of one and a half (in absolute value) at point *A* in the above figure means that she places 1.5 times the value on an extra unit of consumption now as an extra unit of consumption later.

Or, put differently, consumption now is 50% more valuable to her than consumption later.

DISCUSS 11.1: THE CONSEQUENCES OF PURE IMPATIENCE

1. Draw the indifference curves of a person who is always more impatient than Julia in Figure 11.3b for any level of consumption now and consumption later.
2. Draw a set of indifference curves for Julia if she does not experience diminishing marginal returns to consumption. Would she then desire to smooth her consumption?
3. Draw a set of indifference curves for Julia if she does not experience diminishing marginal returns to consumption and has no pure impatience.
4. Based on your answers to the above parts, what must be true of a consumer's preferences for her to want to smooth consumption?

11.4 BORROWING ALLOWS SMOOTHING BY BRINGING CONSUMPTION TO THE PRESENT

How much will Julia borrow? If we combine Figures 11.2 and 11.3a (we assume that the rate of interest is 10%) we will have the answer. As in the other examples of a feasible set and indifference curves, Julia wishes to get to the highest indifference curve, but is limited by her feasible frontier. The highest feasible indifference curve will be the one that is tangent to the feasible frontier shown as point *E* in Figure 11.4.

She chooses to borrow and consume $58 and repay $64 later, leaving her $36 to consume later. Because the slopes of the indifference curve and the feasible frontier are equal at this point (otherwise the curves would cross), we know that the slope of the indifference curve is equal to the slope of the feasible frontier. We define a *person's discount rate*, ρ (economists use the Greek letter *rho*, which rhymes with "toe"), as the slope of the indifference curve minus one, which is a measure of how much Julia values an extra unit of consumption now over an extra unit of consumption later.

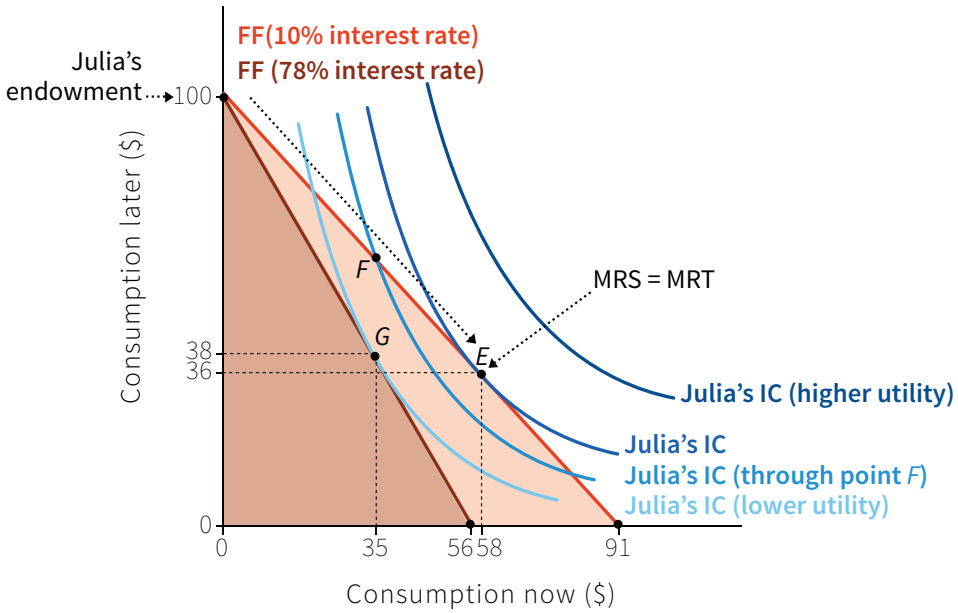


Figure 11.4 *Moving consumption over time by borrowing.*

A PERSON'S DISCOUNT RATE

A *person's discount rate,* ρ, is a measure of a person's impatience, namely how much she values an extra unit of consumption now over an extra unit of consumption later. This is the slope of her indifference curve between consumption now and consumption later, minus one.

Her discount rate depends on two things:

- *Her desire to smooth consumption*: This is affected by the situation she is in—the current distribution of consumption now and later.
- *Her pure impatience as a person*: A person's discount rate is also sometimes referred to as her subjective discount rate because it is based in part on her psychology.

For example, in Figure 11.3b, ρ = 50% at point A because an extra unit of consumption today was worth 1.5 extra units later. This means that Julia borrows just enough so that:

$$\begin{aligned} \textit{slope of the indifference curve} &= MRS \\ &= 1 + \rho \\ &= 1 + r \\ &= MRT \\ &= \textit{slope of the feasible frontier} \end{aligned}$$

And if we subtract 1 from both sides of this equation we have:

$$\textit{discount rate} = \rho = r = \textit{rate of interest}$$

Her discount rate ρ depends on both her desire to smooth consumption and on her degree of pure impatience.

To see why she is doing the best she can by borrowing that much at the given interest rate, suppose that this were not the case and she had chosen some other point (say *F*)—meaning that the slope of the indifference curve at *F* is greater than the slope of the feasible frontier.

This would mean that her discount rate, ρ, exceeded *r*, the rate of interest (or the cost to her of bringing goods from the future to the present. Thus if she found herself at point *F*, she would choose to bring goods forward in time, by borrowing. This eliminates all points on the feasible frontier except *E*, at which her discount rate is equal to the rate of interest.

What happens if the interest rate at which she can borrow increases? As in Figure 11.2, the feasible set gets smaller, and the best she can now do is to borrow less ($35 instead of $58) because it has become more expensive to bring buying power forward in time. The combination of consumption now and consumption later that she chooses at the higher interest rate is shown by point *G* in Figure 11.4.

DISCUSS 11.2: INCOME AND SUBSTITUTION EFFECTS

1. Use Figure 11.4 to show that the difference in current consumption at the lower and higher interest rate (at *E* and *G*), namely $23, is composed of an income effect and a substitution effect. It will be helpful to review income and substitution effects from Unit 3 before doing this.
2. Why do the income and substitution effects work in the same direction in this example?

11.5 LENDING AND STORING: SMOOTHING AND MOVING CONSUMPTION TO TH FUTURE

Now think about Marco, an individual in a different situation from Julia considering a payday loan, or a farmer in Chambar seeking a loan until the harvest. At the same time as Julia is deciding how much to borrow, Marco has some goods or funds worth $100, but does not (yet) anticipate receiving any income later. Julia and Marco will both get $100 eventually, but time creates a difference. Marco's wealth, narrowly defined, is $100. Julia's wealth is zero.

We saw that Julia, earning $100 in the future, wants to borrow. This reflects the situation she is in, which gives her a strong desire to smooth by borrowing, and how impatient she is as a person (her pure impatience). Think about what Julia's indifference curve, passing through her endowment point, might look like. As shown in Figure 11.5 it is very steep. Because she currently has nothing, she has a strong preference for increasing consumption now.

This is called Julia's *reservation indifference curve*, because the curve is made of all of the points at which Julia would be just as well off as at her reservation position, which is her endowment, with no borrowing or lending. (Julia's endowment and reservation indifference curves are similar to those of Angela, the farmer, in Unit 5.)

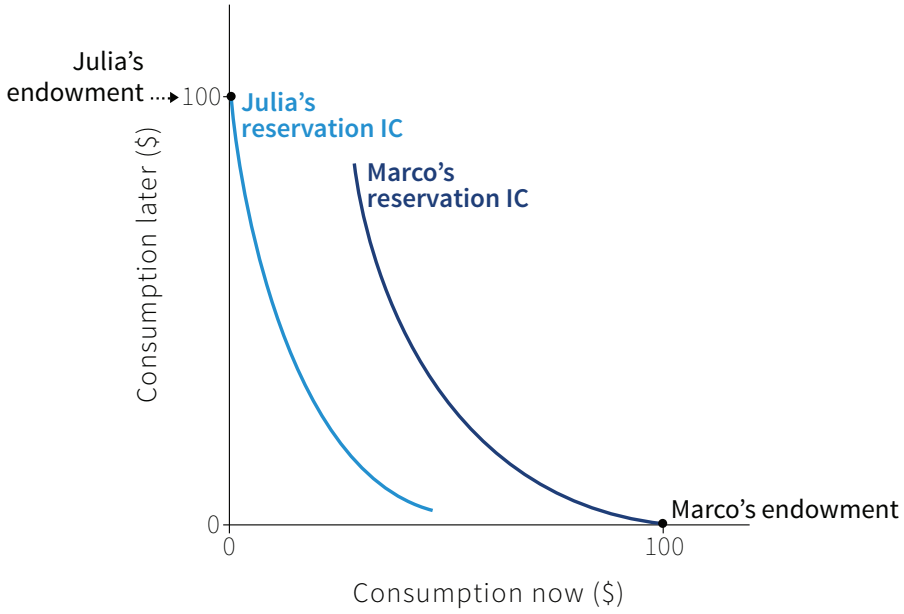


Figure 11.5 *Reservation indifference curves and Julia and Marco's endowments.*

Look at Marco's indifference curve passing thought his endowment point, which is $100 now and nothing later. As shown in Figure 11.5 it is quite flat now, indicating that he does not prefer more consumption now to in the future—in fact he is looking for a way to transfer some of his consumption to the future.

Marco and Julia's indifference curves and hence their pure impatience are similar. Their situations, not their preferences, differentiate them. Julia borrows because she is poor in the present, unlike Marco, and that is why she is impatient: she needs to smooth her consumption.

Marco has $100 worth of grain just harvested, and no debts to pay off. He could consume it all now but, as we have seen, this would probably not be the best he could do in the circumstances:

- We have assumed his income in the future is zero.
- Like Julia, he has diminishing marginal returns to consumption of grain.

In order to smooth, he wishes to move some goods to the future. He could store the grain but, if he did, mice would eat some of it. Mice are a form of depreciation: the grain they eat represents a reduction in Marco's wealth due to the passage of time. Let's suppose that, taking account of the mice, if he consumed nothing at all this period he would have just $80 worth of grain later. This means that the cost of moving grain from the present to the future is 20% per year.

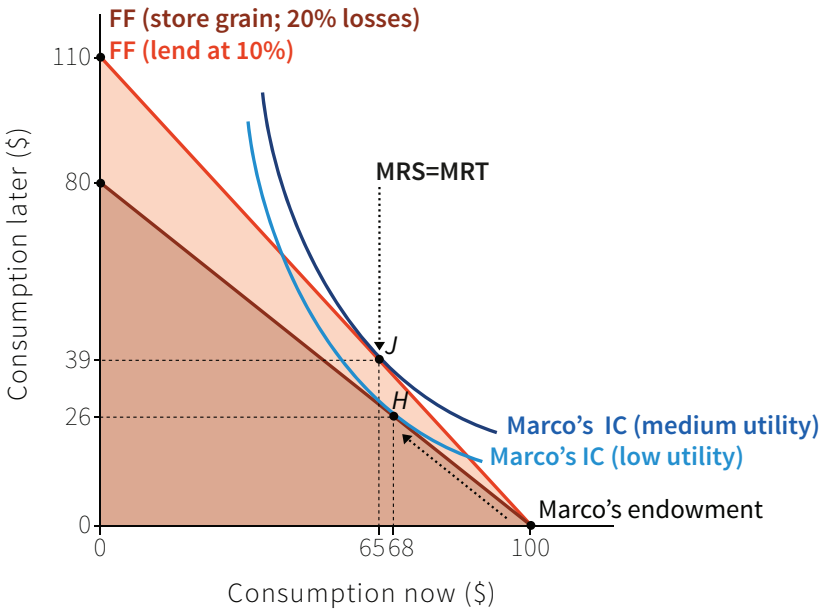


Figure 11.6 *Smoothing consumption by storing and lending.*

In Figure 11.6, we see that Marco's endowment is on the horizontal axis, as he has $100 of grain available now. The dark line shows Marco's feasible frontier using storage, and the dark shaded area shows his feasible set. If this were the only option, and if his indifference curves were as indicated, he would definitely store some of the grain. In Figure 11.6, some part of his feasible frontier lies outside his endowment indifference curve, so he can do better by storing some grain.

But how much? Like Julia he will find the amount of storage that gets him to the highest feasible indifference curve by finding the point of tangency between the indifference curve and the feasible frontier. This is point *H*, so he will eat $68 of the grain now, and consume $26 of it later (mice ate $6 of the grain). At point *H*, Marco has equated his MRS between consumption now and in the future to the cost of moving goods from the present to the future, which is the MRT.

He could avoid the mice by selling the grain and putting $100 under his mattress. His feasible frontier would then be a straight line (not shown) from consumption now of $100, to consumption later of $100. We are assuming that his $100 note will not be stolen and that $100 will purchase the same amount of grain now and later because there is no inflation (we explain inflation, and its effects, in Unit 13). Under these assumptions, storing money under the mattress is definitely better than storing grain among mice.

A better plan, if Marco could find a trustworthy borrower, would be to lend the money. If he did this and could be assured of repayment of $\$(1 + r)$, then he could have $100 \times (1 + r)$ feasible consumption later, or any of the combinations along his new feasible consumption line. The light line in Figure 11.6 shows the feasible frontier when Marco lends at 10%. As you can see from the figure, compared to

storage, or putting the money in his mattress, his feasible set is now expanded by the opportunity to lend money at interest. Marco is now able to reach a higher indifference curve.

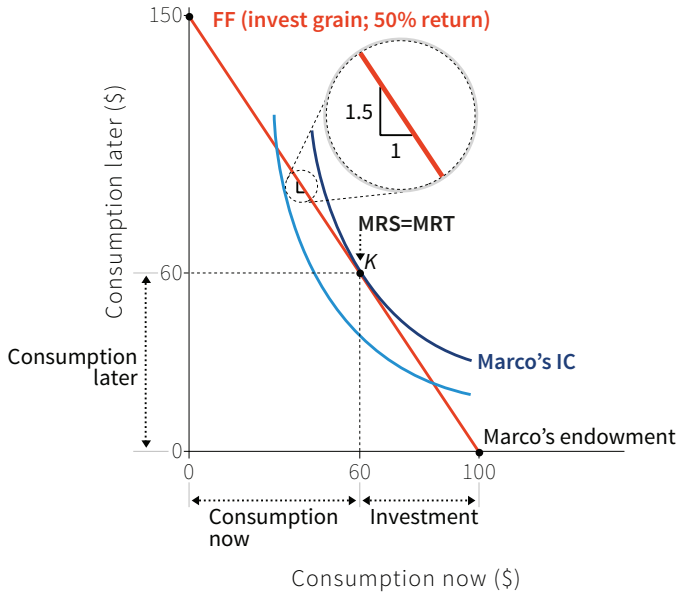
DISCUSS 11.3: AN INCREASE IN THE INTEREST RATE

- Use a diagram like Figure 11.4 to show the income and the substitution effects of an increase in the interest rate for Marco who receives his endowment today.
- How do these effects compare to those for Julia?

11.6 INVESTING: ANOTHER WAY TO MOVE CONSUMPTION TO THE FUTURE

If Marco owns some land he could do even better: he could invest the grain (planting it as seed, feeding it to his draft animals to help him work the fields until harvest). This opportunity to invest will further expand his feasible set. Suppose that if he were to invest all of his grain he could harvest $150 worth of grain later, as shown in Figure 11.7. He has invested $100, harvested $150, and so earned a profit of *$150 - $100 = $50*, or a profit rate (profits divided by the investment required) of *$50/$100 = 50%*. The slope of the red line is -1.5, where the absolute value (1.5) is the marginal rate of transformation of investment into returns, or one plus the rate of return on the investment.

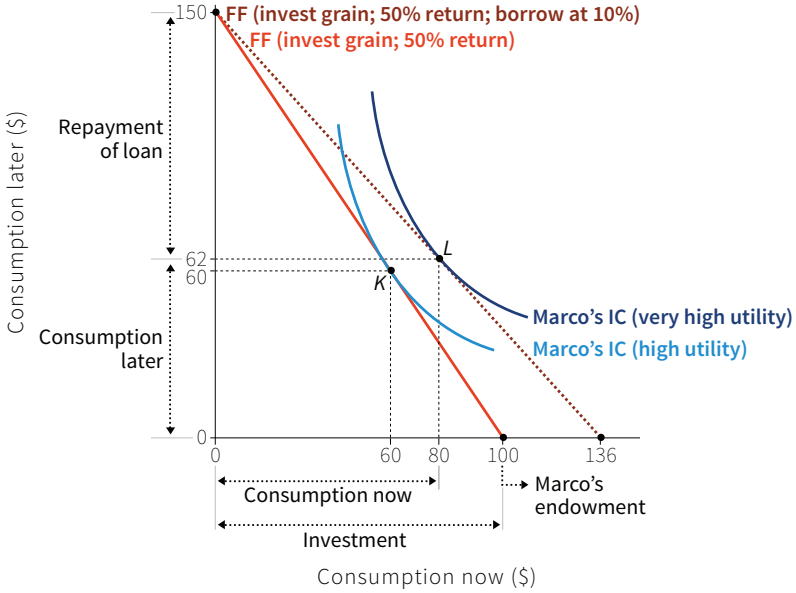
But if he could get a loan at 10% he would quickly see that he would be better off with an entirely new plan: invest everything he has, with a harvest next year of $150, but also borrow now so as to be able to consume more both now and in the future. This "invest it all" plan is shown in Figure 11.8. The plan shifts Marco's feasible frontier out even further, as shown by the dotted red line. Marco ends up consuming at a new point, *L*, with more both now and in the future.



Marco's optimal choice

If Marco were to invest all of his grain he could harvest $150 worth of grain later. The slope of the red line is -1.5, where the absolute value (1.5) is one plus the rate of return on the investment.Marco chooses to consume $60 now and $60 later, as shown by point K. At this point, the feasible frontier is tangent to an indifference curve.

Figure 11.7 *Investing in a high-return project.*



Optimal choice after getting a loan

Marco ends up consuming at point *L*, with $80 now and $62 in the future.

Figure 11.8 *Borrowing to invest in a high-return project.*

Figures 11.9 and 11.10 summarise how the "invest it all and borrow" plan works compared to the other options.

PLAN (POINT IN FIGURE 11.10)	Rate of return or interest	Consumption now	Consumption later	Investment	Ranking by utility (or combined consumption)
STORAGE (H)	-20% (a loss)	$68	$26	None	4th ($94)
LENDING ONLY (J)	10%	$65	$39	None	3rd ($104)
INVESTMENT ONLY (K)	50%	$60	$60	$40	2nd ($120)
INVESTMENT AND BORROWING (L)	50% (investment), -10% (lending)	$80	$62	$100	1st ($142)

Figure 11.9 *Storage, lending, investment and borrowing provide Marco with many feasible sets.*

These feasible sets are shown in Figure 11.10.

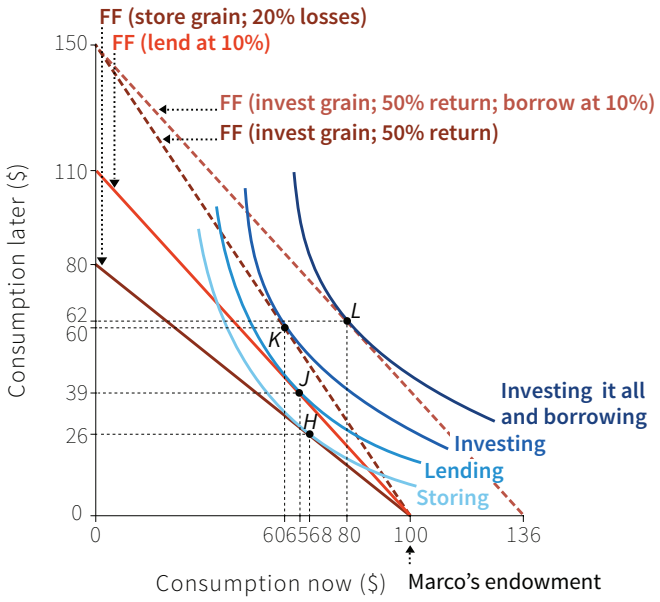


Figure 11.10 *Options for the individual (Marco) who starts with assets.*

Let's return to how Marco differs from Julia. Compare the feasible sets of Julia shown in Figure 11.4 and of Marco, whose options are shown in Figure 11.10.

Two differences between Marco and Julia explain the disparity in their outcomes.

- *Marco starts with an asset while Julia starts with nothing*: Julia has the prospect of a similar asset later, but this puts the two on opposite sides of the credit market. If Marco wants to smooth his consumption so that he will have something to consume later, he has money to lend and to invest. Without assets Julia necessarily has to borrow to be able to consume now.

  For Marco, interest will be income, and a higher interest rate pushes the feasible frontier outwards, enlarging the set of feasible consumption opportunities. For Julia, interest is a cost, and a higher rate of interest pushes the feasible frontier inward, shrinking the feasible set.
- *Marco and Julia face different interest rates*: The less obvious difference is that if Marco (after investing his entire asset at a 50% return) wants to move his buying power forward in time, he borrows against his future income at a rate of 10%. Julia, lacking assets like the poor farmers in Chambar, would most likely have no alternative but to borrow at the higher rate of 78%. The paradox is that *Marco can borrow at a low interest rate because he does not need to borrow*.

To summarise, borrowing, lending, storing and investing are ways of moving claims on goods forward (to the present) or backwards (to the future) in time.

People engage in these activities because:

- *They prefer a smooth path of consumption*: They have diminishing marginal returns to consumption.
- *They might increase their consumption in both periods by these activities* (as Marco did by investing and borrowing).

People differ in which of these activities they engage (some borrowing, some lending) because:

- *They have differences in their situation:* For example, having an income now or later, which will affect their discount rates and their opportunities.
- *They differ in their level of pure impatience.*

DISCUSS 11.4: LIFETIME INCOME

Consider an individual's income over his or her lifetime from leaving school to retirement. Explain how an individual may move from a situation like Julia's to one like Marco's in the course of a lifetime (assume that their pure impatience remains unchanged over their lifetime).

11.7 ASSETS, LIABILITIES AND NET WORTH

We will see that a person's wealth is an important aspect of their situation in the process of borrowing, lending, and investing, and that those with more wealth like Marco have opportunities not open to those with less wealth, like Julia. Balance sheets are an essential tool for understanding how wealth changes when an individual, or a firm, borrows and lends.

A balance sheet summarises what the household or firm owns, and what it owes to others. What you own (including what you are owed by others) is called your assets, and what you owe others is called your liabilities. (To be liable means to be responsible for something, in this case to make good your debts to others.) The difference between your assets and your liabilities is called your net worth. The relationship between assets, liabilities and net worth is shown in Figure 11.11.

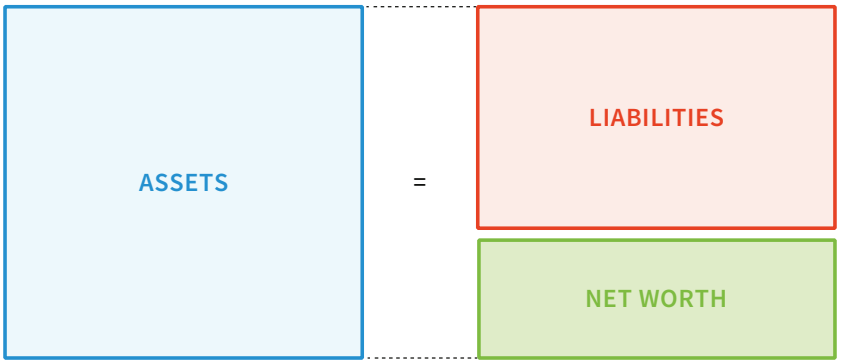


Figure 11.11 *A balance sheet.*

When the components of an equation are such that by definition, the left-hand side is equal to the right-hand side, it is called an accounting identity, or identity for short. The balance sheet identity states:

$$assets \equiv liabilities + net\ worth$$

To understand the concept of net worth, which is what makes the left- and right-hand sides balance by definition, we can turn the identity around by subtracting liabilities from both sides so that:

$$\begin{aligned} net\ worth &\equiv assets - liabilities \\ &\equiv what\ the\ household\ owns\ or\ is\ owed\ to\ it \\ &\quad - what\ the\ household\ owes\ to\ others \end{aligned}$$

In the bathtub model the water in the bathtub represented wealth, which is the same thing as net worth. As we saw, net worth or wealth increases with income. It declines with consumption and depreciation.

But your wealth or net worth does not change when you lend or borrow. This is because a loan creates both an asset and a liability on your balance sheet: if you borrow money you receive cash as an asset, while the debt is an equal liability.

Julia started off with neither assets nor liabilities and a net worth of zero, but on the strength of her expected future income she borrowed $58 now when the interest rate was 10% (point *E* in Figure 11.4). At this time her asset is the $58 in cash that she is holding, while her liability is the loan that she has to pay back later. We record the value of the loan as $58 now, since that is what she received for getting into debt (its value rises to $64 later only once interest has been added). This is why taking out the loan has no effect on her current net worth—the liability and the asset are equal to one another, so her net worth remains zero. In Figure 11.12 this is recorded in her balance sheet under the heading "Now (before consuming)".

Now (before consuming)

JULIA'S ASSETS		JULIA'S LIABILITIES	
Cash	$58	Loan	$58
		Net worth = $58 - $58	0

Now (after consuming)

JULIA'S ASSETS		JULIA'S LIABILITIES	
Cash	0	Loan	$58
		Net worth	-$58

Later (before consuming)

JULIA'S ASSETS		JULIA'S LIABILITIES	
Cash	$100	Loan	$64
		Net worth = $100 - $36	$36

Later (after consuming)

JULIA'S ASSETS		JULIA'S LIABILITIES	
Cash	$64	Loan	$64
		Net worth	0

Figure 11.12 *Julia's balance sheets.*

She then consumes the $58—it flows out through the bathtub drain, to use our earlier analogy. Since she still has the $58 liability, her net worth falls to -$58. This is recorded in Figure 11.12 in her balance sheet under the heading "Now (after consuming)".

Later, she receives income of $100 (an inflow to the bathtub). Also, because of accumulated interest, the value of her loan has risen to $64. So her net worth becomes $100 - $64 = $36. Again, we suppose that she then consumes the $36, leaving her with $64 in cash to pay off her debt of $64. At this point her net worth falls back to zero. The corresponding balance sheets are also shown in Figure 11.12.

11.8 BANKS, MONEY AND THE CENTRAL BANK

Among the moneylenders in Chambar, the profitability of the loans business depends on:

- The cost of their borrowing
- The default rate on the loans they extended to farmers
- The interest rate they set

The closure of Irish banks for six months revealed how money can be created in an economy and how it depends on trust.

These case studies and the two-period model provide much of what we need in order to understand the role of the financial system in the economy. But we need to introduce two more actors on the economic stage: banks and the central bank. When we focus on banks we will greatly simplify the financial system. We do this because banks have a special relationship with the government.

We saw in Unit 2 that a capitalist economy works through the success of some firms, but also through the failure and elimination of other, low-performing ones. Firms go out of business every day. The government steps in to save them only in very unusual cases. The same logic doesn't apply to banks. Some banks seem to live forever: the Monte dei Paschi bank in Siena, Italy is more than 500 years old.

Governments rescue banks when they are considered too big, or too important for the continued functioning of the economy, to be allowed to fail. Unlike the failure of a firm, a banking crisis can bring down the financial system as a whole and threaten the livelihoods of people throughout the economy. This is why we concentrate on banks rather than other financial sector firms.

A bank is a firm that makes profits by lending and borrowing. Banks lend to households and firms on different terms from the terms on which they borrow from them. It may seem odd to say that banks borrow from households but that is exactly what they do: they take deposits (money placed in the bank), which they promise to repay when requested. The interest they pay on these deposits is lower than the interest they charge when they make loans, and this allows them to make profits.

To explain this process, we first have to explore in more detail the concept of money.

We saw that anything that is accepted as payment can be counted as money. But money in this sense is different from *base money*, also called *legal tender* or *high-powered money*. Unlike cheques, legal tender has to be accepted as payment by law. It comprises cash (notes and coins), and accounts held by commercial banks at the *central bank*.

Most of what we count as money is not legal tender issued by the central bank: instead, it is created by commercial banks when they make loans. We explain using bank balance sheets.

CENTRAL BANK

- The central bank is the only bank that can create legal tender.
- The government usually owns the central bank.
- It is also the banker for the commercial banks: commercial banks have accounts at the central bank, holding legal tender.

Suppose that Marco has $100 in cash and he puts it in a bank account in Abacus Bank. Abacus Bank will put the cash in a vault, or it will deposit the cash in its account at the central bank. Abacus Bank's balance sheet gains $100 of base money as an asset, and the liability of a bank balance of $100 payable on demand to Marco, as shown in Figure 11.13a.

ABACUS BANK'S ASSETS		ABACUS BANK'S LIABILITIES	
Base money	$100	Payable on demand to Marco	$100

Figure 11.13a *Marco deposits $100 in Abacus Bank.*

Marco wants to pay $20 to his local grocer, Gino, in return for groceries so he instructs Abacus Bank to transfer the money to Gino's account in Bonus Bank. (He could do this by writing a cheque to Gino, or paying Gino using a debit card.) This is shown on the balance sheets of the two banks in Figure 11.13b: Abacus Bank's assets and liabilities both go down by $20, while Bonus Bank's assets increase by the $20 of base money, and its liabilities increase by $20 payable on demand to Gino.

ABACUS BANK'S ASSETS		ABACUS BANK'S LIABILITIES	
Base money	$80	Payable on demand to Marco	$80

BONUS BANK'S ASSETS		BONUS BANK'S LIABILITIES	
Base money	$20	Payable on demand to Gino	$20

Figure 11.13b *Marco pays $20 to Gino.*

This illustrates the payment services provided by banks. So far we have just considered transactions using base money, or legal tender. We now show how banks create money in the process of making loans.

Suppose that Gino borrows $100 from Bonus Bank. Bonus Bank lends him the money by crediting his bank account with $100, so he is now owed $120. But he owes a debt of $100 to the bank. So Bonus Bank's balance sheet has expanded: its assets have grown by the $100 it is owed by Gino, and its liabilities have grown by the $100 it has credited to his bank account, shown in Figure 11.13c.

BONUS BANK'S ASSETS		BONUS BANK'S LIABILITIES	
Base money	$20	Payable on demand to Gino	$120
Bank loan	$100		
Total	$120		

Figure 11.13c *Bonus Bank gives Gino a loan of $100.*

Bonus Bank has now expanded the money supply: Gino can write cheques up to $120, so in this sense the money supply has grown by $100—even though base money has not grown. The money created by his bank is called *bank money.*

Base money remains essential, however, partly because customers sometimes take out cash, but also because when Gino wants to spend his loan, his bank has to transfer base money. Suppose Gino employs Marco to work in his shop, and pays him $10. Then Bonus Bank has to transfer $10 of base money from Gino's bank account to Marco's bank account in Abacus Bank. This transaction is shown in Figure 11.13d.

ABACUS BANK'S ASSETS		ABACUS BANK'S LIABILITIES	
Base money	$90	Payable on demand to Marco	$90

BONUS BANK'S ASSETS		BONUS BANK'S LIABILITIES	
Base money	$10	Payable on demand to Gino	$110
Bank loan	$100		
Total	$110		

Figure 11.13d *Gino pays Marco $10.*

In practice, banks make many transactions to one another in a given day, most cancelling each other out, and they settle up at the end of each day. So at the end of each day each bank will transfer, or receive, the net amount of transactions they have made. This means they do not need to have available the legal tender to cover all transactions or demand for cash.

The total "money" in the banking system has grown, as Figure 11.13e shows.

ASSETS OF ABACUS BANK AND BONUS BANK		LIABILITIES OF ABACUS BANK AND BONUS BANK	
Base money	$100	Payable on demand	$200
Bank loan	$100		
Total	$200		

Figure 11.13e *The total money in the banking system has grown.*

Creating money may sound like an easy way to make profits, *but the money they create is a liability, not an asset, because it has to be paid on demand to the borrower*. It is the corresponding loan that is an asset for the bank. Banks make profits out of this process by charging interest on the loans. So if Bonus Bank lends Gino the $100 at an interest rate of 10%, then next year the loan—the bank's asset—has grown in value to $110, exceeding the value of the original liability. Since net worth is equal to the value of assets minus the value of liabilities, this allows banks to create positive net worth.

Base money *plus* bank money is called *broad money*.

The ratio of base money to broad money varies across countries and over time. For example, before the financial crisis base money comprised about 3-4% of broad money in the UK, 6-8% in South Africa, and 8-10% in China.

TYPES OF MONEY

Money is a medium of exchange used to purchase goods or services. It can take the form of bank notes, cheques, credit, or whatever else one purchases things with.

- *Base money* (also known as legal tender): Cash and the reserves of commercial banks held in their accounts at the central bank.
- *Bank money:* Money created by commercial banks when they extend credit to firms and households.
- *Broad money:* The stock of money in the economy, which is the sum of base and bank money.

By taking deposits and making loans banks provide the economy with the service of *maturity transformation*. Bank depositors—either individuals or firms—can withdraw their money from the bank without notice. But when banks lend they give a fixed date—which, in the case of a mortgage loan for a house purchase, may be 30 years in the future—on which the loan will be repaid. They cannot require the borrower to repay sooner, thus allowing those receiving bank loans to engage in long-term planning. This is called maturity transformation because the length of a loan is termed its maturity, so the bank is engaging in short-term borrowing and long-term lending. (It is also called liquidity transformation: The lenders' deposits are liquid—free to flow out of the bank on demand—but bank loans to borrowers may be considered the opposite of liquid—frozen—requiring time before they can be converted to a flow of income that the bank can use.)

While maturity transformation is an essential service in any economy, it also exposes the bank to a new form of risk (called *liquidity risk*), beyond the possibility that its loans will not be repaid (called *default risk*).

Banks make money by lending much more than they hold in legal tender, because they count on depositors not to need their funds all at the same time. The risk it faces is that depositors can decide they want to withdraw money instantaneously, and the money won't be there. In Figure 11.13e the banking system owed $200, but only held $100 of base money. If all customers demanded their money at once, the banks would not be able to repay. This is called a *bank run*. If there's a run, the bank is in trouble, as this article explains. Liquidity risk is a cause of bank failures.

11.9 THE CENTRAL BANK, THE MONEY MARKET AND INTEREST RATES

Commercial banks make money out of banking services and loans. But for that they need to be able to make transactions, for which they need base money. There is no automatic relationship between the amount of base money they require, and the amount of lending they do. Rather, they need whatever amount of base money will cover the net transactions they have to make on a daily basis. The price of borrowing base money is the *short-term interest rate*.

Suppose in the above example that Gino wants to pay $50 to Marco (and there are no other transactions that day). Gino's bank, Bonus Bank, doesn't have enough base money to make the transfer to Abacus Bank, as we can see from its balance sheet in Figure 11.13f.

BONUS BANK'S ASSETS		BONUS BANK'S LIABILITIES	
Base money	$20		
Bank loan	$100	Payable on demand to Gino	$120
Total	$120		

Figure 11.13f *Bonus bank does not have enough base money to pay $50 to Abacus Bank.*

So Bonus Bank has to borrow $30 of base money to make the payment. Banks borrow from one another in the money markets since, at any moment, some banks will have excess cash and others not enough. They could also try to induce someone to deposit additional money in another bank account, but deposits also have costs due to interest payments, marketing, and maintaining bank branches. Thus cash deposits are only one part of bank financing.

But what determines the price of borrowing in the money market (the interest rate)? We can think in terms of supply and demand:

- *The demand for base money* depends on how many transactions commercial banks have to make.
- *The supply of base money* is simply a decision by the central bank.

Since the central bank controls the supply of base money it can also decide the interest rate. The central bank intervenes in the money market by saying it will lend whatever quantity of base money is demanded at the interest rate i that it chooses.

Banks in the money market will respect that price: no bank will borrow at a higher rate, or lend at a lower rate, since they can borrow at rate i from the central bank. This i is also called the *base rate, official rate* or *policy rate*.

The base rate applies to banks that borrow base money from each other, and from the central bank. But it matters in the rest of the economy because of its knock-on effect on other interest rates in the economy. The average interest rate charged by commercial banks to firms and households is called the *bank lending rate*. This rate will typically be above the policy interest rate, to ensure that banks make profits. (It will also be higher for borrowers perceived as risky by the bank, as we saw earlier.) The difference between the bank lending rate and the base rate is the markup or *spread* on commercial lending.

In the UK, for example, the policy interest rate set by the Bank of England was 0.5% in 2014, but few banks would lend at less than 3%. In emerging economies the gap can be quite large, owing to the uncertain economic environment. In Brazil, for instance, in 2014 the central bank policy rate was 11% but the bank lending rate was 32%.

The central bank does not control this mark-up, but generally the bank lending rate goes up and down with the base rate—just as other firms typically vary their prices according to their costs.

Figure 11.14 greatly simplifies the financial system. We show savers facing just two choices: to deposit money in a bank current account, which we assume pays no interest, or buy government bonds in the money market. The interest rate on government bonds is called the yield. Go to our Einstein section for an explanation of these bonds, and why the yield on government bonds is close to the policy interest rate. We also give an explanation of what are called *present value* calculations, which are essential for you to understand how assets like bonds are priced.

We have now seen how the central bank sets the policy interest and how this affects the lending interest rate. But why should a central bank do this at all? This is really two questions:

1. *How does the lending rate affect spending in the economy*? We will answer this question in Section 11.11.
2. *Why does the central bank wish to affect spending by changing the interest rate*? (As we suggest in the top box in Figure 11.14?) We will answer this much larger question in Units 12-14. In those units we explain fluctuations in employment and inflation in the economy as a whole, and reasons why central banks are frequently given responsibility for moderating those fluctuations by changing the interest rate.

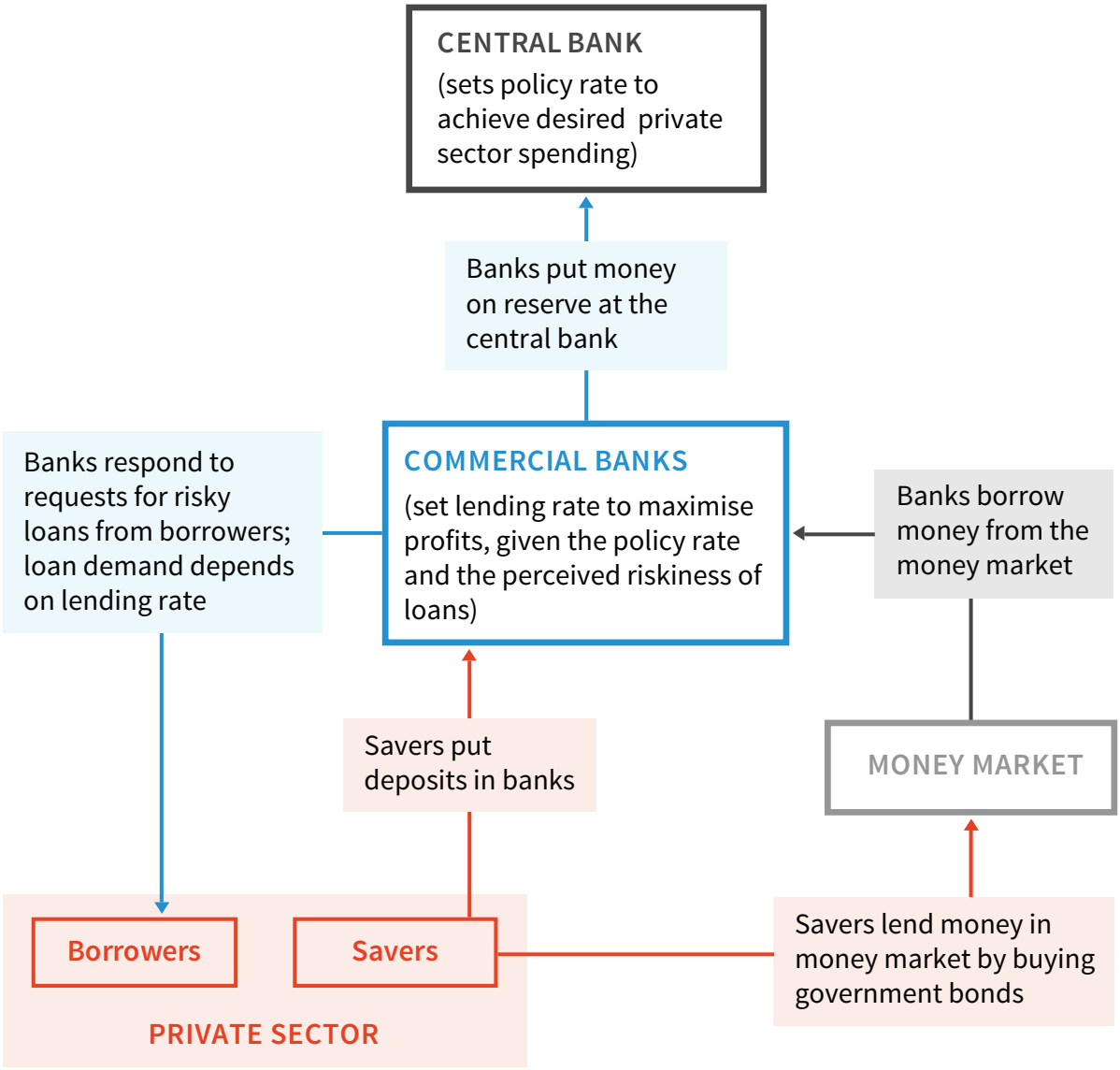


Figure 11.14 *Banks, the central bank, borrowers and savers.*

Source: *Adapted from Figure 5.12 in Chapter 5 of Carlin, Wendy, and David Soskice. 2014. Macroeconomics: Institutions, Instability, and the Financial System. Oxford: Oxford University Press.*

DISCUSS 11.5: THE POLICY INTEREST RATE AND MONEY SUPPLY

1. Draw a diagram with the amount of broad money supplied by the central bank and the banking system on the horizontal axis, and the policy interest rate on the vertical axis. Draw a downward-sloping demand curve. Suppose the central bank sets the policy interest rate at 5%. Indicate the amount of money supplied at that interest rate.
2. Suppose the demand for money goes up (suggest why this might occur) and indicate what happens to the amount of money supplied by the central bank and the banking system.
3. Draw the curve for the money supply in this model.

Recall that a variable in a model is called endogenous when its value is the outcome of the relationships in the model. A variable is called exogenous when its value comes from outside the model.

4. Is the policy interest rate exogenous or endogenous?
5. Is the money supply exogenous or endogenous? Justify your answers.

DISCUSS 11.6: INTEREST RATE MARK-UPS

Use the web sites of two central banks of your choice to collect data on the monthly policy interest rate and the mortgage interest rate between 2000 and the most recent year available.

1. Plot the data.
2. How does the banking mark-up (interest rate margin) compare between the two countries?
3. Do banking mark-ups change over time? Suggest possible reasons for what you observe.

11.10 THE BUSINESS OF BANKING AND BANK BALANCE SHEETS

To understand the business of banking in more detail, we look at a bank's costs and revenues:

- *The bank's operational costs:* These include the administration costs of making loans. For example, the salaries of loan officers who evaluate loan applications, the costs of renting and maintaining a network of branches and call centres used to supply banking services.
- *The bank's interest costs:* They must pay interest on their liabilities, including deposits and other borrowing.
- *The bank's revenue:* This is the interest and repayment of the loan it receives from its customers.
- *The bank's expected return:* This is the return on the loans it provides, taking into account that not all customers will repay their loans.

Like moneylenders, if the risk banks take when making loans, the default rate, is higher, then there will be a larger gap (or spread or markup) between the interest rate they charge on the loans they make, and the cost of their borrowing.

The profitability of the business depends on the difference between the cost of borrowing and the return to lending, taking account of the default rate and the operational costs of screening the loans and running the bank.

A good way to understand a bank is to look at its entire balance sheet, which summarises its core business of lending and borrowing. Banks borrow and lend to make profits:

- *Bank borrowing is on the liabilities side:* Deposits, secured and unsecured borrowing are recorded as liabilities.
- *Bank lending is on the assets side:* This is financed by the borrowed funds.

Figure 11.15 shows the simplified balance sheet of a commercial bank.

ASSETS (owned by the bank or owed to it)		% of balance sheet	LIABILITIES (what the bank owes households, firms and other banks)		% of balance sheet
1. Cash and reserve balances at central bank	Owned by the bank: immediately accessible funds	2	1. Deposits	Owned by households and firms	50
2. Financial assets, some of which (government bonds) may be used as collateral for borrowing	Owned by the bank	30	2. Secured borrowing (collateral provided)	Including borrowing from other banks via the money market	30
3. Loans to other banks	Via the money market	11	3. Unsecured borrowing (no collateral provided)		16
4. Loans to households and firms (e.g. mortgages)		55			
5. Fixed assets such as buildings and equipment	Owned by the bank	2			
Total assets		100	Total liabilities		96
			4. Net worth = Total assets - total liabilities = equity		4

Figure 11.15 *A simplified bank balance sheet*

Source: *Adapted from Figure 5.9 in Chapter 5 of Carlin, Wendy, and David Soskice. 2014. Macroeconomics: Institutions, Instability, and the Financial System. Oxford: Oxford University Press.*

As we saw above:

$$\begin{aligned} \textit{net worth} &\equiv \textit{assets} - \textit{liabilities} \\ &\equiv \textit{what the household owns or is owed to it} \\ &- \textit{what the household owes to others} \end{aligned}$$

Another way of saying this is that the net worth of a firm like a bank is equal to what is owed to the shareholders or owners. This explains why net worth is on the liabilities side of the balance sheet. If the value of the bank's assets is less than the value of what the bank owes others, then its net worth is negative: it is *insolvent*.

The asset side of the bank balance sheet

1. Item 1 on the balance sheet is the cash it holds, plus the bank's balance in its account at the central bank, which is called its reserve balances. Cash and reserves at the central bank are the bank's readily accessible, or liquid, funds. This is base money and amounts to a tiny fraction of the bank's balance sheet—just 2% in this example of a typical contemporary bank. As we saw above, money created by the central bank is a very small proportion of the broad money that circulates in the economy.
2. Banks own financial assets. One use of these assets is as collateral for the bank's borrowing in the money market. As we discussed above, they borrow to replenish their cash balances (item 1) when depositors withdraw (or transfer) more funds than the bank has available.
3. A bank will also have loans to other banks on its balance sheet.
4. The bank's lending activities are the largest item on the asset side. The loans made by the bank to households and firms make up 55% of the balance sheet in Figure 11.15. This is the bank's core business. Some of this will be secured lending. A loan is secured if the borrower has provided collateral. In the case of housing loans, called mortgages, the value of the house is the collateral. Other bank loans are unsecured like overdrafts, credit card balances and consumer loans.
5. Bank assets such as buildings and equipment will be recorded on the asset side of the balance sheet.

The liability side of the bank balance sheet

On the liability side there are three forms of bank borrowing, shown in lines 1, 2 and 3 in Figure 11.15.

1. The most important one is bank deposits, making up 50% of the bank's balance sheet in this example. The bank owes these to households and firms. As part of its profit-maximisation decision, the bank makes a judgement about the likely demand by depositors to withdraw their deposits. Across the banking system withdrawals and deposits occur continuously, and when the cross-bank transactions are cleared, most cancel each other out. Any bank must ensure that it has cash and reserves at the central bank to meet the demand by depositors for funds. Holding cash and reserves for this purpose has an opportunity cost, because they could instead be lent out in the money market in order to earn interest; banks hold the minimum prudent balances of cash and reserves.

The second and third entries on the liabilities side of the balance sheet are what the bank has borrowed from households, firms and other banks in the money market.

2. Some of this is secured borrowing, for which the bank provides collateral using its financial assets (which appear on the left-hand side of the balance sheet in item 2).
3. Some borrowing is unsecured. As we shall see, the cost of these funds is set by the central bank's policy interest rate.

The final item on the balance sheet is the bank's *net worth*. This is the bank's *equity*. For a typical bank, its equity is only a few per cent of its balance sheet. The bank is a very debt-heavy company.

Figure 11.16 shows the simplified balance sheet of Barclays bank (just before the financial crisis) and Figure 11.17 shows the simplified balance sheet of a company from the nonfinancial sector, Honda.

ASSETS		LIABILITIES	
1. Cash and reserve balances at the central bank	7,345	1. Deposits	336,316
2. Wholesale reverse repo lending	174,090	2. Wholesale repo borrowing secured with collateral	136,956
3. Loans (e.g. mortgages)	313,226	3. Unsecured borrowing	111,137
4. Fixed assets (e.g. buildings, equipment)	2,492	4. Trading portfolio liabilities	71,874
5. Trading portfolio assets	177,867	5. Derivative financial instruments	140,697
6. Derivative financial instruments	138,353	6. Other liabilities	172,417
7. Other assets	183,414		
Total assets	996,787	Total liabilities	969,397
		NET WORTH	
		Equity	27,390
Memorandum item: Leverage (Total assets/Net worth)		996,787/27,390 = 36.4	

Figure 11.16 *Barclays Bank balance sheet in 2006 (£m).*

Source: *Barclays Bank. 2006. 'Barclays Bank PLC Annual Report.' Also presented as Figure 5.10 in Chapter 5 of Carlin, Wendy, and David Soskice. 2014. Macroeconomics: Institutions, Instability, and the Financial System. Oxford: Oxford University Press.*

ASSETS		LIABILITIES	
1. Current Assets	5,323,053	1. Current liabilities	4,096,685
2. Finance subsidiaries-receivables net	2,788,135	2. Long-term debt	2,710,845
3. Investments	668,790	3. Other liabilities	1,630,085
4. Property on operation leases	1,843,132		
5. Property, plant and equipment	2,399,530		
6. Other assets	612,717		
Total assets	13,635,357	Total liabilities	8,437,615
		NET WORTH	
		Equity	5,197,742
Memorandum item: Leverage as defined for banks: (Total assets/Net worth)		13,635,357/5,197,742 = 2.62	
Memorandum item: Leverage as normally defined for non-banks: (Total Liabilities/ Total assets)		8,437,615/13,635,357 = 61.9%	

Figure 11.17 *Honda Motor Company balance sheet in 2013 (¥m).*

Source: *Honda Motor Co. 2013. 'Annual Report.'*

Current assets refers to cash, inventories and other short-term assets. Current liabilities refer to short-term debts and other pending payments.

A way of describing the reliance of a company on debt is to refer to its *leverage* or gearing.

Unfortunately the term leverage is defined differently for financial and nonfinancial companies (both definitions are shown in Figures 11.16 and 11.17). We calculate the leverage for Barclays and Honda using the definition used for banks: total assets divided by net worth. Barclays' total assets are 36 times their net worth. This means that given the size of its liabilities (its debt), a very small change in the value of its assets (1/36 ≈ 3%) would be enough to wipe out its net worth and make the bank insolvent. By contrast, using the same definition, we see that Honda's leverage is less than three. Compared to Barclays, Honda's equity is far higher in relation to its

assets. Another way to say this is that Honda finances its assets by a mixture of debt (62%) and equity (38%), whereas Barclays finances its assets with 97% debt and 3% equity.

11.11 THE CENTRAL BANK'S POLICY RATE CAN AFFECT SPENDING

Households and firms borrow to spend: the more it costs to borrow (equivalently, the higher the interest rate), the less they spend. This allows the central bank to influence the amount of spending in the economy, which then affects firms' decisions about how many people to employ and what prices to set. In this way the central bank can affect the level of unemployment and inflation (rising prices), as we shall see in detail in Units 12 to 14.

To see the effect of a lower interest rate on consumption spending, we return to Julia, who has no wealth, but expects to receive $100 one year from now. We can see from the left-hand panel of Figure 11.18 that at the moneylender's interest rate of 78% she borrowed in order to spend $35 now, at G, but at the interest rate of 10% she would borrow and spend $58 now, at E. The right-hand panel of the figure traces out Julia's consumption spending now as the interest rate falls. The points G and E in the right-hand panel correspond to those in the left-hand panel. The downward-sloping line is Julia's demand for loans, which also shows her expenditure now. So lowering the interest rate will increase consumption spending.

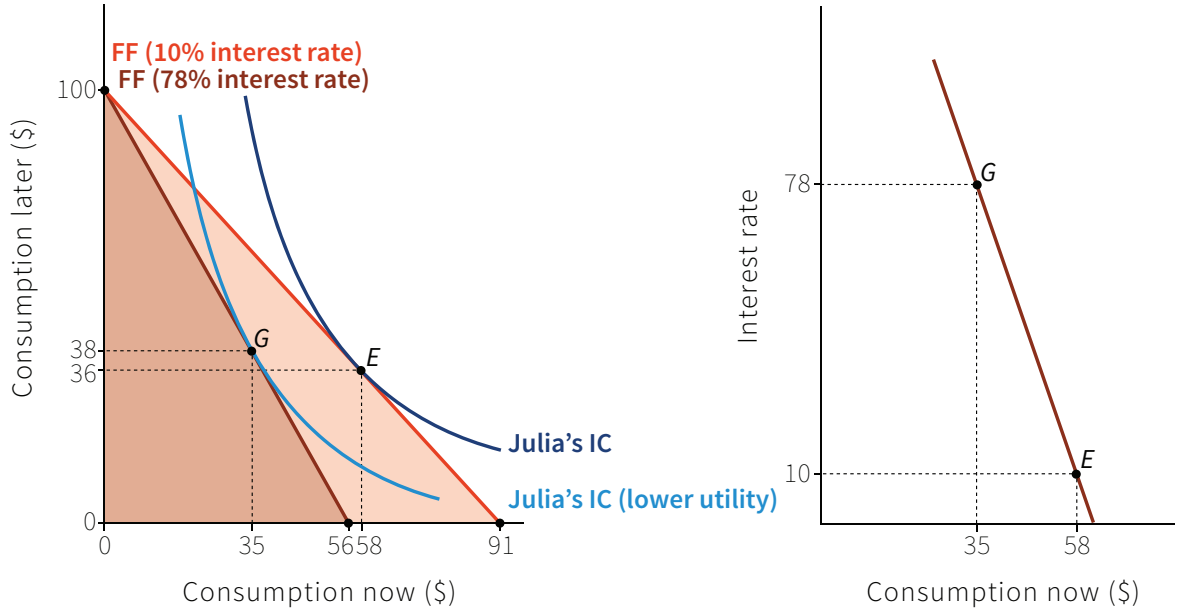


Figure 11.18 *Interest rates and consumption spending.*

In many rich countries, when people borrow it is most often to purchase a car or a home (mortgages for housing are less common in countries where financial markets are less developed). Loans for this purpose are readily available to people even of limited wealth because unlike a loan to purchase food or daily consumption items, the thing purchased—the car or the house—can be signed over to the bank as collateral insuring the bank against default risk. For this reason, an important channel for the effects of the interest rate on domestic spending in many rich economies is through its effect on home purchases and consumer durables such as automobiles. Central banks can help to moderate ups and downs in spending on housing and consumer durables, and hence smooth out the fluctuations in the whole economy.

DISCUSS 11.7: INTEREST RATES AND CONSUMPTION SPENDING

Think about the income and substitution effects of a rise in the interest rate analysed in Discuss 11.3. Comment on whether a rise in the interest rate would be expected to reduce consumption expenditure in an economy, where a proportion of households are like Julia's, a proportion is like Marco's and the rest are neither lenders nor borrowers.

11.12 MONEY AND CREDIT MARKET CONSTRAINTS: A PRINCIPAL-AGENT PROBLEM

Lending is risky. A loan is made now and has to be repaid in the future. Between now and then unanticipated events beyond the control of the borrower can occur. An earthquake, disease, the outbreak of war or the obsolescence of the skill you have invested in using your student loan are unavoidable risks, and will mean the loan may not be repaid. The rate of interest set by a bank or a moneylender will be greater, the greater is the risk of default due to unavoidable events.

But lenders face another problem. The lender cannot be sure that a borrower will exert enough effort to make the project succeed. Nor can the lender be as good a judge of whether the project is likely to succeed as the borrower. Both of these problems arise from the difference between the information the borrower and the lender have about the borrower's project and actions.

This creates a conflict of interest. If the project doesn't succeed because the borrower made too little effort, or because it just wasn't a good project, the lender loses money. If the borrower were using only her own money, it is likely that she would have been more conscientious or maybe not engaged in the project at all.

The relationship between the lender and the borrower is termed a *principal-agent problem*. The lender is the "principal" and the borrower is the "agent". The principal would like his agent to act in a certain way—work hard to make the project succeed, only seek a loan for a promising project—but cannot ensure that this will happen because the principal does not have and cannot get the information that would allow him to write these requirements into a contract that can be enforced.

PRINCIPAL-AGENT RELATIONSHIP

This relationship exists when one party (the principal) would like another party (the agent) to act in some way or have some attribute...

- ... that is in the interest of the principal but not the agent
- ... that cannot be enforced or guaranteed in a binding contract

The principal-agent problem between borrower and lender is similar to the "somebody else's money" problem discussed in Unit 6: the manager of a firm (the agent) is making decisions about the use of the funds supplied by the firm's investors (the principals), but they are not in a position to require him to act in a way that maximises their wealth, rather than pursuing his own objectives.

You have seen another principal-agent problem, also in Unit 6. The manager of the firm (now, the principal) cannot contractually enforce his objective that employees (the agents) work hard and well. In all three cases—manager and employee, owners and manager, lender and borrower—the fundamental problem is that there is a conflict of interest and the available information does not allow the principal to write an enforceable contract that ensures that he will get what he wants from the interaction.

Principal-agent problems always have two features:

- There is a conflict of interest between the principal and the agent.
- This conflict is about some *hidden action or attribute* of the agent that cannot be enforced or guaranteed in a binding contract.

Figure 11.19 compares two principal-agent problems.

	Actors	Conflict of interest over	Enforceable contract covers	Left out of contract (or unenforceable)	Market failure
Labour market (Unit 6)	Employer Employee	Wages, work (quality & amount)	Wages, time, conditions	Work (quality and amount), duration of employment	Effort under-provided; unemployment
Credit market (Unit 11)	Lender Borrower	Interest rate, conduct of project (effort, prudence)	Interest rate	Effort, prudence, repayment	Too much risk, credit constraints

Figure 11.19 *The credit market and the labour market as principal-agent problems.*

You know from Unit 10 that, when this is the case, the result will be a market failure, that is, an outcome that is not Pareto efficient. We will see that this is the case in the credit market.

One response of the lender to the conflict of interest is to require the borrower to put some of her wealth into the project (this is called equity). The more of the borrower's own wealth is invested in the project the more closely aligned their interests are with those of the lender. Whether farmers in Chambar or car buyers in New York, often the lender requires the borrower to set aside property that will be transferred to the lender if the loan is not repaid (this is called *collateral*).

Equity or collateral reduces the conflict of interest between the borrower and the lender. The reason is that when the borrower has some of her money (either equity or collateral) at stake:

- *She has a greater interest in working hard:* She will try harder to make prudent business decisions to ensure the project's success.
- *It is a signal to the lender:* It signals that she thinks that the project is of sufficient quality to succeed.

But there is a hitch: if the borrower had been wealthy, she could either use her wealth as collateral and as equity in the project, or she could have been on the other side of the market, lending money. Typically the reason why the borrower needs a loan is that she is not wealthy. As a result, she may be unable to provide enough equity or collateral to sufficiently reduce the conflict of interest and hence the risk faced by the lender, and the lender refuses to offer a loan.

This is called *credit rationing*: those with less wealth borrow on unfavourable terms compared with those with more wealth, or are refused loans entirely.

Borrowers whose limited wealth makes it impossible to get a loan at any interest rate are termed *credit-excluded*. Those who borrow, but on unfavourable terms, are termed *credit-constrained*. Both are sometimes said to be wealth-constrained, meaning that their wealth limits their credit market opportunities. Adam Smith had credit rationing in mind when he wrote:

"Money, says the proverb, makes money. When you have got a little it is often easy to get more. The great difficulty is to get that little."
—Adam Smith, 'Of the Profits of Stock.' *In An Inquiry into the Nature and Causes of the Wealth of Nations, 1776.*

The relationship between wealth and credit is summarised in Figure 11.20.

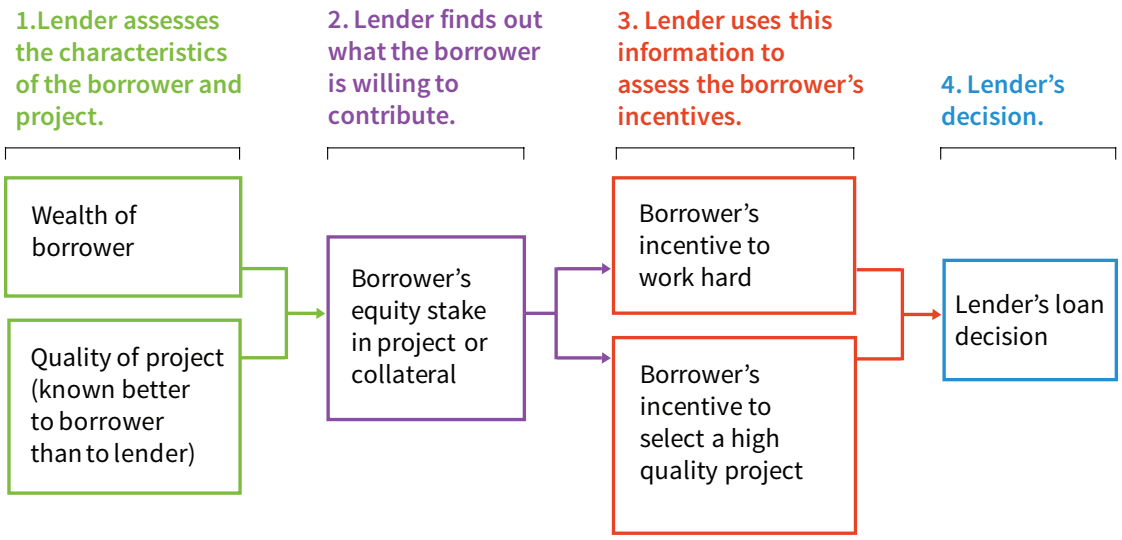


Figure 11.20 *Wealth, project quality, and credit.*

The exclusion of those without wealth from credit markets or their borrowing on unfavourable terms is evident in these facts:

- One-fifth of US families in a survey had their request for credit rejected by a financial institution; the assets of these credit-constrained families were 63% lower than the unconstrained families. "Discouraged borrowers" (those who did not apply for a loan because they expected to be rejected) had even lower wealth than the rejected applicants.
- Credit card borrowing limits are often increased automatically. If borrowing increases in response to these exogenous changes in the borrowing limit, we can conclude that the individual was credit-constrained. The authors of this study concluded that something like two-thirds of US families are credit-constrained or excluded.
- An inheritance of $10,000 doubles a typical British youth's likelihood of setting up in business; and inheritance leads the self-employed to increase the scale of their operations considerably.

- A 10% rise in value of housing assets that could be used as collateral to secure loans in the United Kingdom increases the number of startup businesses by 5%.
- Asset-poor people in the US frequently take out short term "payday loans". In the state of Illinois, the typical short-term borrower is a low-income woman in her mid-thirties ($24,104 annual income), living in rental housing, borrowing between $100 and $200 and paying an average annual rate of interest of 486%.
- Poor and middle-income Indian farmers could substantially raise their incomes if they did not confront credit constraints: not only do they underinvest in productive assets generally, but the assets they hold are biased towards those they can sell in times of need (bullocks) and against highly profitable equipment (irrigation pumps), which have little resale value.

11.13 MONEY AND CREDIT MARKET FAILURES AND SUCCESSES

In modern economies, as we have just shown, money is created and its supply determined by private banks through their lending activity, and by the central bank through issuing legal tender. Governments and large organisations regulated by governments (the banks) are central to this process, and are today essential for it to work well.

When the Irish banks shut down, private citizens created money by accepting cheques that could not be cleared in the banking system, but which were trusted as if they were bank notes. Money and credit are so useful that they tend to crop up even when central banks and the banking system don't provide them. Archaeologists have discovered evidence of lending, and the use of money to denominate debts and to facilitate exchange, long before banks or governments existed.

This occurred because of the mutual gains made possible when a group of people come to accept a particular medium of exchange and develop sufficient trust among one another so that borrowing and lending, and sellers accepting the medium of exchange, occur. The mutual gains associated with borrowing and lending can be seen from the example of Julia and Marco, if we assume for a moment that Julia borrows money from Marco rather than from a bank. Both would be better off when Julia wished to bring some of her future buying power to the present, and Marco wanted to shift some of his present buying power to the future. An important part of their story is that Julia paid interest as a result, and Marco received it. So, while there were mutual gains, there also were conflicts of interest.

Money and credit have also been essential to the process of innovation throughout history. They have allowed a person with the idea of a new technology, or way of organising a business, to borrow the funds for the necessary development, build

prototypes, and profit from the innovation. Lending to entrepreneurs by large industrial banks, for example, fuelled rapid growth of the German economy in the late 19th century.

But sometimes credit markets fail in the same way that Unit 10's market for bananas failed when it rewarded the plantation owners' use of an environmentally destructive pesticide. The failure occurred because plantations did not pay the full cost of their use of Weevokil. The reason for the credit market failure is the same as the rest of the market failures in Figure 10.8 (see Section 10.10): when people make decisions about borrowing and lending, whether bankers or borrowers, they do not take into account the effect of their actions on others.

In credit markets, as we have seen (Figure 11.20) the bank is more likely to offer a loan to a borrower who has enough wealth to put up collateral or to invest her own equity in the project. Some borrowers are entirely excluded from the credit market (as we saw) and others are offered loans only in small amounts, at very high rates of interest, or for exceptionally good projects.

To put it another way, lenders are willing to trade off project quality to get a borrower who has more equity or more collateral. Sometimes a high-quality project from a poor would-be borrower is not funded by the lender, while a rich individual with a middling project gets a loan, as shown in Figure 11.21.

PROJECT QUALITY	RICH	POOR
HIGH	Loan granted	No loan
INTERMEDIATE	Loan granted	No loan
LOW	No loan	No loan

Figure 11.21 *Project quality and wealth of borrower.*

The result is another market failure that often goes by the name credit market exclusion or credit constraints. Figure 11.22 is an additional line in the summary table of market failures in the previous unit, Figure 10.11.

The decision	How it affects others	Cost or benefit	Market failure (misallocation of resources)	Possible remedies	Terms applied to this type of market failure
Misrepresentation of the quality of a project; insufficient prudence or effort is its success	Imposes risk of non-repayment on lender	Cost	Too little credit is extended to poor people with good projects	Redistribute wealth; common responsibility for repayment of loans (Grameen Bank)	Credit market exclusion, credit constraints

Figure 11.22 *Credit market failure.*

DISCUSS 11.8: THE LENDER'S DILEMMA

Read the paper *The Microfinance Promise*. The Grameen Bank in Bangladesh makes loans available to groups of individuals who together apply for individual loans, with the proviso that the loans to the group members will be renewed in the future if (but only if) each member has repaid the loan on schedule.

Think how such an arrangement will help address the "lender's dilemma" above, specifically the borrower's decision about what to spend the money on, and how hard she will work to make sure that repayment is possible.

Morduch, Jonathan. 1999. 'The Microfinance Promise.'*Journal of Economic Literature* 37 (4): 1569–1614.

Credit market failures occur also for an additional reason. When the bank makes a loan, it takes account of the possibility that it may not be repaid: if the interest rate it can charge is sufficiently high, even quite risky loans—like the payday loans mentioned at the outset—may be a good bet. The bank also worries about what might happen to its profits should most of its borrowers be unable to pay, as would happen for example if it had extended mortgages for home purchases during a boom in housing prices, and then the bubble burst. The bank could fail.

If the owners of the bank would bear all of the costs of a bankruptcy, then they would take extraordinary efforts to avoid it. But the owners are unlikely to bear the full costs, for two reasons:

- *The bank typically will have borrowed from other banks:* Just like the farmer borrowing to plant his crop, the bank owners will know that some of the costs of bankruptcy will be borne by banks that will not be repaid.
- *"Too big to fail":* If the bank that may fail is sufficiently important in the economy, then the prospect of its failure is likely to provoke a bailout of the bank by the government, subsidising it with tax revenue.

So again the bank owners know that others (tax payers or other banks) will bear some of the costs of their risk-taking. So they take more risks than they would were they to bear all of the costs of their actions. Excess risk-taking by banks and borrowers is—like environmental spillovers—a negative external effect leading to a market failure.

Those who may get stuck with the risk-taker's losses try to protect themselves. Governments seek to regulate the banking system, limiting bank leverage so that banks would theoretically have sufficient resources to repay their debts.

The banks also protect themselves, as we have seen, by requiring that a borrower commit equity or collateral to the project. This partially addresses the conflict of interest between the bank owners and the borrowers, reducing the extent of the market failure resulting from the borrowers' too-risky choices.

But this also creates a third market failure. When banks try to protect themselves, it creates credit market constraints, which mean that the economy's resources will be devoted to implementing projects of lesser quality than would otherwise be the case.

11.14 CONCLUSION

Banks are not the most popular or trusted institutions. In the US, for example, 72% of people expressed "a great deal" or "quite a lot" of confidence in the military in 2015—very similar to the level a decade earlier. By contrast, in 2015, only 28% expressed the same confidence in banks, down from 49% a decade earlier. Surveys show the public in Germany, Spain and many other countries hold their banks in low esteem. This has particularly been the case since the financial crisis of 2008.

It is sometimes said that rich people lend on terms that make them rich, while poor people borrow on terms that make them poor. Our example of Julia and Marco made it clear that one's view of the interest rate—as a cost for Julia and as a source of income for Marco—depends on one's wealth. People with limited wealth are credit-constrained, which limits their ability to profit from the investment opportunities that are open to those with more assets.

It is also true that, in determining the rate of interest at which an individual will borrow, the lender often has superior bargaining power, and so can set a rate so as to capture most of the mutual gains from the transaction.

But do banks and the financial system make some people poor and other people rich? To answer this question, compare banks to other profit-making firms. Both are owned by wealthy people, who profit from the business they do with poorer people. Moreover, they often transact on terms (rates of interest, wages) such that the lack of wealth of borrowers and employees is perpetuated.

But even those who dislike banks do not think that the less wealthy would be better off in their absence, any more than that the less wealthy would benefit if firms ceased to employ labour. Banks, credit and money are essential to a modern economy—including to the economic opportunities of the less well off—because they provide opportunities for mutual gains that occur when people can benefit by moving their buying power from one time period to another, either borrowing (moving it to the present) or lending (the opposite).

CONCEPTS INTRODUCED IN UNIT 11

Before you move on, review these definitions:

- *Money, Broad money, Base money, Bank money*
- *Wealth*
- *Income*
- *Diminishing marginal returns to consumption*
- *A person's discount rate*
- *Pure impatience*
- *Collateral*
- *Balance sheet, Assets, Liabilities, Net worth, Equity, Solvency*
- *Leverage*
- *Credit-constrained; Credit-excluded*
- Central bank's *policy interest rate*
- *Principal-agent problem*

In Unit 19 we will consider policies to improve the capacity of the financial system to contribute to the economy. Such policies, like any economic policy or institution, should begin by asking: how can the financial system be organised so as both to allow all possible mutual gains to be realised and to result in outcomes that are fair?

Our understanding of the labour market from Unit 6, the credit market from this unit and the process of innovation from Unit 2 provide the basis for our understanding of how the economy considered as a whole works, which is our next subject.

DISCUSS 11.9: UNPOPULAR BANKS

Why do you think that banks tend to be more unpopular than other profit-making firms (Honda or Microsoft, for example)?

DISCUSS 11.10: LIMITS ON LENDING

Many countries have policies to limit how much interest a moneylender can charge on a loan.

1. Do you think these limits are a good idea?
2. Who benefits from the laws and who loses?
3. What are likely to be the long-term effects of such laws?
4. Contrast this approach to helping the poor gain access to loans with the Grameen Bank in Discuss 11.8.

Key points in Unit 11

Money

Money allows purchasing power to be transferred among people so that they can exchange goods and services.

The rate of interest on borrowed money

This is the price of moving spending from the future to the present; the rate of return on an investment project is the price one receives by delaying spending to a future date.

How much to borrow or invest?

For a person with a given level of wealth and expected future income, the amount borrowed or invested depends on:

- The interest rate
- The rate of return on investment
- The person's discount rate at their endowment point

The poor are credit-constrained

Those with little wealth to post as collateral or to invest in a project are often excluded from the credit market or are credit-constrained (able to borrow only at high interest rates, or small amounts, or for unusually good projects).

Banks produce money by supplying credit

They set interest rates so as to make profits; the central bank produces money by issuing legal tender.

The central bank affects spending

By altering the policy interest rate and thereby changing the lending rate that banks choose to set on their loans, the central bank affects the amount of spending by households and firms.

Financial markets may fail

Like other markets, financial markets allow mutually advantageous exchanges to occur, but also sometimes fail in doing this and may also contribute to unfair outcomes and to economic instability.

11.15 EINSTEIN

Present value (PV)

Assets like shares in companies, bank loans, or bonds typically provide a stream of income in the future. Since these assets are bought and sold, we have to ask the question: how do we value a stream of future payments? The answer is the *present value* (PV) of the expected future income.

To make this calculation we have to assume that people in the market to buy and sell assets have the capability to save and borrow at a certain interest rate. So imagine you face an interest rate of 6% and are offered a financial contract that says you will be paid €100 in one year's time. That contract is an asset. How much would you be willing to pay for it today?

You would not pay €100 today for the contract, because if you had €100 today you could put it in the bank and get €106 in a year's time—which would be better than buying the asset.

Imagine you are offered the asset for €90 today. Now you will want to buy it, because you could borrow €90 today from the bank at 6%, and in a year's time you would pay back €95.40 while you receive €100 from the asset, making a profit of €4.60.

The break-even price, €P, for this contract would make you indifferent between buying the contract and not buying it. It has to be equal to whatever amount of money would give you €100 in a year's time if you put it in the bank today. With an interest rate of 6%, that amount is:

$$P = \frac{100}{1+6\%} = \frac{100}{1.06} = €94.34$$

€94.34 today is worth the same to you as €100 in a year's time because if you put €94.34 in the bank then it would be worth €100 in a year. Equivalently, if you borrowed €94.34 today from the bank to buy the asset you would have to pay back €100 in a year's time, exactly offsetting the €100 the asset gives us.

We say that the income next year is *discounted* by the interest rate: a positive interest rate makes it worth less than income today.

The same logic applies further in the future, where we allow for interest compounding over time. If you receive €100 in t years time, then today its value to you is:

$$P_t = \frac{100}{1+6^t}$$

Now suppose an asset gives a payment each year for T years, paying X_t in year t, starting next year in year 1. Then each payment X_t has to be discounted according to how far in the future it is. So with an interest rate of r the PV of this asset is:

$$PV = \frac{X_1}{(1+r)^1} + \frac{X_2}{(1+r)^2} + \ldots + \frac{X_T}{(1+r)^T}$$

The present value of these payments obviously depends on the amounts of the payments themselves. But it also depends on the interest rate: if the interest rate increases, then the PV will decline, because future payments are discounted (their PV reduced) by more.

Net present value (NPV)

This logic applies to any asset that provides income in the future. So if a firm is considering whether or not to make an investment, they have to compare the cost of making the investment with the present value of the profits they expect it to provide in the future. In this context we consider the net present value (NPV), which takes into account the cost of making the investment as well as the expected profits. If the cost is C and the present value of the expected profits is PV, then the NPV of making the investment is:

$$NPV = PV - C$$

If this is positive then the investment is worth making, because the expected profits are worth more than the cost (and vice versa).

Bond prices and yields

A bond is a particular kind of financial asset, where the bond issuer promises to pay a set amount over time to the bondholder. Issuing or selling a bond is equivalent to borrowing, because the bond issuer receives cash today and promises to repay in the future. Conversely, a bond buyer is a lender or saver, because they give up cash today, expecting to be repaid in the future. Both governments and firms borrow by issuing bonds.

Bonds typically last a predetermined amount of time, called the *maturity* of the bond, and provide two forms of payment: the *face value F*, which is an amount paid at the end of the period (when the bond *matures*), and a fixed payment every period (for example, every year or every quarter) until then. In the past bonds were physical pieces of paper and when one of the fixed payments was redeemed, a coupon was clipped from the bond. For this reason the fixed payments are called *coupons* and we label them C.

As we saw in the calculation of PV, the amount that a lender will be willing to pay for a bond will be its present value, which depends on the bond's face value, the series of coupon payments, and also on the interest rate. No one will buy a bond for more than

its present value because they would be better off putting their money in the bank. No one will sell a bond for less than its present value, because they would be better off borrowing from the bank. So:

price of bond = discounted present value of coupons
+ discounted present value of the face value when it matures

Or, for a bond with a maturity of T years:

$$P = \underbrace{\frac{C}{(1+r)^1} + \frac{C}{(1+r)^2} + \ldots + \frac{C}{(1+r)^T}}_{\textit{coupons}} + \underbrace{\frac{F}{(1+r)^T}}_{\textit{face value}}$$

An important characteristic of a bond is its yield. This is the implied return that the buyer gets on their money when they buy the bond at its market price. We calculate the yield using an equation just like the PV equation. The yield y will solve the following:

$$P = \underbrace{\frac{C}{(1+y)^1} + \frac{C}{(1+y)^2} + \ldots + \frac{C}{(1+y)^T}}_{\textit{coupons}} + \underbrace{\frac{F}{(1+y)^T}}_{\textit{face value}}$$

If the interest rate stays constant, as we have assumed, then the yield will be the same as that interest rate. But in fact we cannot be sure how interest rates are going to change over time. In contrast, we know the price of a bond, its coupon payments and its face value, so we can always calculate a bond's yield. Buying a bond with yield y is equivalent to saving your money at the guaranteed constant interest rate of $r = y$.

Since a saver (a lender) can choose between buying a government bond, lending the money in the money market, or putting it into a bank account, the yield on the government bond will be very close to the rate of interest in the money market. If it weren't, money would be switched very quickly from one asset to the other until the rates of return were equalised. This is an example of *arbitrage*.

Let's take a numerical example: a government bond with a face value of €100, yearly coupon of €5, and a maturity of four years. The interest rate in the money market is 3%, and we use this to discount the cash flows we receive.

So the price of this bond is given by:

$$\frac{5}{(1.03)^1} + \frac{5}{(1.03)^2} + \frac{5}{(1.03)^3} + \frac{5}{(1.03)^4} + \frac{100}{(1.03)^4}$$

$$= 4.854 + 4.713 + 4.576 + 4.442 + 88.849$$

$$= €107.434$$

We would be willing to pay at most €107.43 for this bond today, even though it generates €120 of revenue over four years. The yield is equal to the interest rate of 3%. If the central bank raises the policy interest rate, then this will reduce the market price of the bond, increasing the yield in line with the interest rate.

11.16 READ MORE

Bibliography

1. Aleem, Irfan. 1990. 'Imperfect Information, Screening, and the Costs of Informal Lending: A Study of a Rural Credit Market in Pakistan.' *The World Bank Economic Review* 4 (3): 329–49.
2. Barclays Bank. 2006. 'Barclays Bank PLC Annual Report.'
3. Bowles, Samuel. 2006. *Microeconomics: Behavior, Institutions, and Evolution (the Roundtable Series in Behavioral Economics)*. Princeton, NJ: Princeton University Press.
4. Carlin, Wendy, and David Soskice. 2014. *Macroeconomics: Institutions, Instability, and the Financial System*. Oxford: Oxford University Press.
5. Farrell, Sean Patrick. 2014. 'The Remote Repo Man.' *The New York Times*, September 24.
6. Gross, David, and Nicholas Souleles. 2002. 'Do Liquidity Constraints and Interest Rates Matter for Consumer Behavior? Evidence from Credit Card Data.' *The Quarterly Journal of Economics* 117 (1): 149–85.
7. Honda Motor Co. 2013. 'Annual Report.'
8. Johnson, Simon, and James Kwak. 2011. *13 Bankers: The Wall Street Takeover and the next Financial Meltdown*. New York: Knopf Doubleday Publishing Group.
9. Martin, Felix. 2013. *Money: The Unauthorised Biography*. London: The Bodley Head.
10. McLeay, Michael, Amar Radia, and Ryland Thomas. 2014. 'Money in the Modern Economy: An Introduction.' *Quarterly Bulletin of the Bank of England* Q1.
11. Morduch, Jonathan. 1999. 'The Microfinance Promise.' *Journal of Economic Literature* 37 (4): 1569–1614.
12. Murphy, Antoin E. 1978. 'Money in an Economy without Banks: The Case of Ireland.' *The Manchester School* 46 (1): 41–50.
13. Riffkin, Rebecca, and Frank Newport. 2015. 'Confidence in US Institutions Still below Historical Norms.' Gallup Inc. November 17.
14. Silver-Greenberg, Jessica. 2014. 'New York Prosecutors Charge Payday Loan Firms with Usury.' *New York Times DealBook*, August 11.

15. Smith, Adam. (1776) 2003. 'Of the Profits of Stock.' In *An Inquiry into the Nature and Causes of the Wealth of Nations*, by Adam Smith. New York, NY: Random House Publishing Group.
16. Spaliara, Marina-Eliza. 2009. 'Do Financial Factors Affect the Capital–labour Ratio? Evidence from UK Firm-Level Data.' *Journal of Banking & Finance* 33 (10): 1932–47.
17. The Economist. 2012. 'The Fear Factor.' June 2.
18. 'The Myth of Barter.' 2012. In *Debt: The First 5,000 Years*, by David Graeber. Brooklyn, NY: Melville House Publishing.

12

ECONOMIC FLUCTUATIONS AND UNEMPLOYMENT

HOW ECONOMIES FLUCTUATE BETWEEN BOOMS AND RECESSIONS AS THEY ARE CONTINUOUSLY HIT BY GOOD AND BAD SHOCKS

- Fluctuations in the total output of a nation (GDP) affect unemployment, and unemployment is a serious hardship for people
- Economists measure the size of the economy using the national accounts: these measures track economic fluctuations and growth
- Households respond to shocks by saving, borrowing and sharing to smooth their consumption of goods and services
- Due to limits on people's ability to borrow (credit constraints) and their weakness of will, these strategies are not sufficient to eliminate shocks to their consumption
- Investment spending by firms (on capital goods) and households (on new housing) fluctuates more than consumption

Institute for New Economic Thinking
UCL SciencesPo Azim Premji University
FRIENDS PROVIDENT Foundation

Losing your job hurts. It causes stress. Following the global financial crisis in 2008, unemployment went up, as did the number of searches for antistress medication on Google. By plotting the increase in intensity of search against the increase in the unemployment rate in the different states of the US in Figure 12.1, we see that in the states where unemployment went up most between 2007 and 2010, searches for antistress medication went up by more too. This suggests that higher unemployment is related to higher stress: we say the two are correlated.

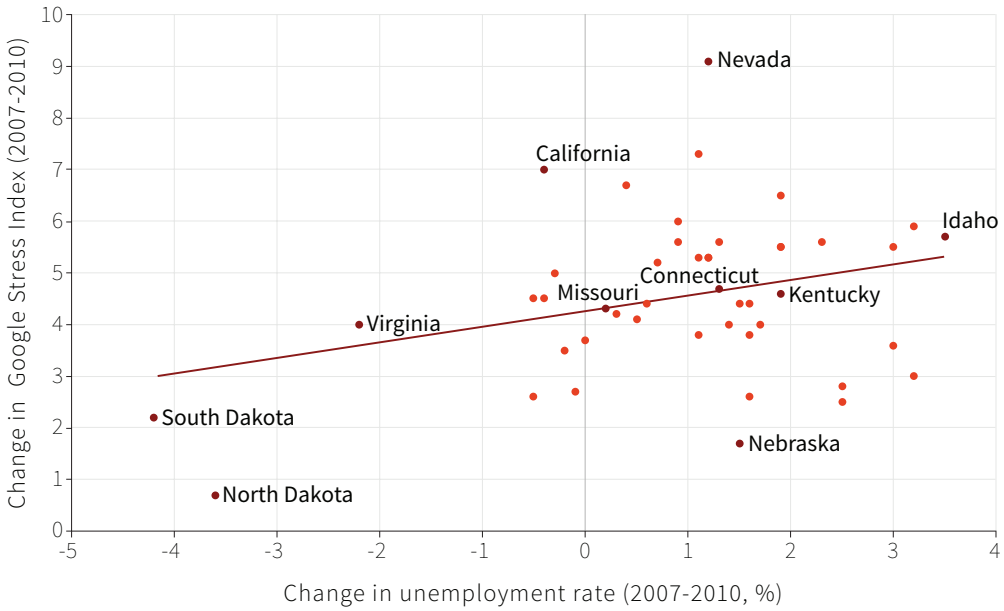


Figure 12.1 *Changes in unemployment and wellbeing during the financial crisis: Evidence from the US states (2007-2010).*

Source: *Algan, Yann, Elizabeth Beasley, Florian Guyot, and Fabrice Murtin. 2014. 'Big Data Measures of Human Well-Being: Evidence from a Google Stress Index on US States.' Sciences Po Working Paper.*

The upward-sloping line summarises the data by finding the line that best fits the scatter of points. This is called a line of best fit or a *linear regression* line. When a line of best fit is upward-sloping it says that higher values of the variable on the horizontal axis (in this case the rise in unemployment) are associated with higher values of the variable on the vertical axis (in this case, the increase in Google searches for antistress medication).

Many kinds of evidence show that being unemployed or fearing unemployment is a major source of unhappiness for people. It ranks alongside major disease and divorce as a stressful life event.

Economists have estimated that becoming unemployed produces more unhappiness than is measured by the loss of earnings from being out of work. Economists Andrew Clark and Andrew Oswald have measured how important life events affect how happy people claim to be when they are asked. In 2002 they calculated that the average

British person would need to be compensated £15,000 ($22,500) per month after losing their job to be as happy as they were when they were employed. This is above the loss of earnings (which at the time were, on average, £2,000 per month).

The compensation needed to restore wellbeing is an enormous amount, much greater than the monetary loss associated with a spell of unemployment. The reason is that unemployment dramatically reduces self-esteem and leads to a much greater reduction in happiness. As we saw in Unit 1, wellbeing depends on more than just income.

CORRELATION MAY NOT BE CAUSATION

Can we draw the conclusion from the data in Figure 12.1 that higher unemployment *causes* higher stress? Maybe we have it the wrong way round, and actually Google searches cause unemployment. Economists call this *reverse causality*. We can rule this out because it is unlikely that an individual Google search on the side-effects of antidepressants could cause an increase in unemployment at the state level. Yet there are other possible explanations for the pattern.

A natural disaster like Hurricane Katrina in the state of Louisiana in the US in 2005 could have triggered an increase in both stress and in job destruction. This is an example where a third factor—in this case, the weather—might account for the finding that searches for antidepressants and unemployment are positively correlated: meaning that as one rises, the other one rises as well. It warns us to be careful in drawing a conclusion from an observed correlation that one trend must have caused the other.

To establish what is causing what, economists devise experiments (like those in Unit 4) or exploit natural experiments (like the comparison of East and West Germany in Unit 1 or the estimate of the size of employment rents in Unit 6).

In the question below we show you a tool that you can use to examine your ideas about how the overall wellbeing in a country can be compared with wellbeing in other countries. What is your recipe for a better life in your country? How important do you think unemployment is? Do other things matter more or just as much—for example, good education, clean air, a high level of trust among citizens, high income, or not too much inequality?

In this unit we learn about why economies go through upswings, during which unemployment falls, and downswings, during which it rises. We focus on the total spending—by households, firms, the government and people outside the home economy—on the goods and services produced by people employed in the home economy.

DISCUSS 12.1: THE OECD BETTER LIFE INDEX

Follow the link to learn about the OECD Better Life Index, which lets you design a measure of the quality of life in a country by deciding how much weight to put on each component of the index.

1. Should a better life index include the following elements: income, housing, jobs, community, education, environment, civic engagement, health, life satisfaction, safety and work-life balance? Justify your answer.
2. Use the Better Life Index tool to create your own better life index for the country where you are living. How does this country score on the topics that are important to you? Rank the countries in the database using your own newly created better life index, and compare it with a ranking based exclusively on income. Choose two countries with contrasting rankings according to the two indices and briefly suggest why this is the case.

12.1 GROWTH AND FLUCTUATIONS

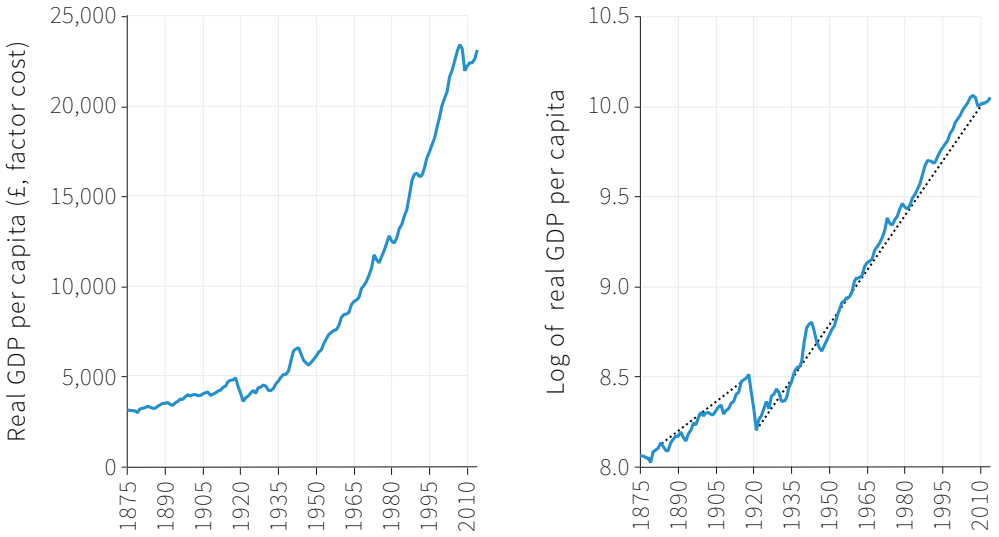
Economies in which the capitalist revolution has taken place have grown over the long run, as illustrated in the hockey stick charts for GDP per capita in Unit 1.

But growth has not been smooth. Figure 12.2 shows the case of the British economy, for which data over a long period is available. The first chart shows GDP per head of the population from 1875. This is part of the hockey stick graph from Unit 1. The chart next to it shows the same data but plots the natural logarithm ("log") of GDP per capita. This is the same as the ratio scale that we used in Unit 1.

See this unit's Einstein to explore the relationship between plotting the log of a variable and the use of a ratio scale on the vertical axis.

By looking at the graph in levels of GDP per capita in the left-hand panel of Figure 12.2, it is hard to tell whether the economy was growing at a steady pace, accelerating or decelerating over time. Transforming the data into natural logs in the right-hand panel allows us to answer the question about the pace of growth more easily. For example, focusing on the period after the first world war, a straight line from 1921 to 2014 fits the data well. For a graph in which the vertical access is measured using the log of GDP per capita the slope of the line (shown in dashed black) represents

the average annual growth rate of the series. Immediately we notice that growth was steady from 1921 to 2014 (with a little uptick during the second world war). Using a spreadsheet or calculator we can work out that the average growth rate over the period was 2.0% per annum. In contrast, the average growth rate was only 0.9% from 1875 to 1914. You can see that a line drawn through the log series from 1875 to 1914 would be flatter than the line from 1921.



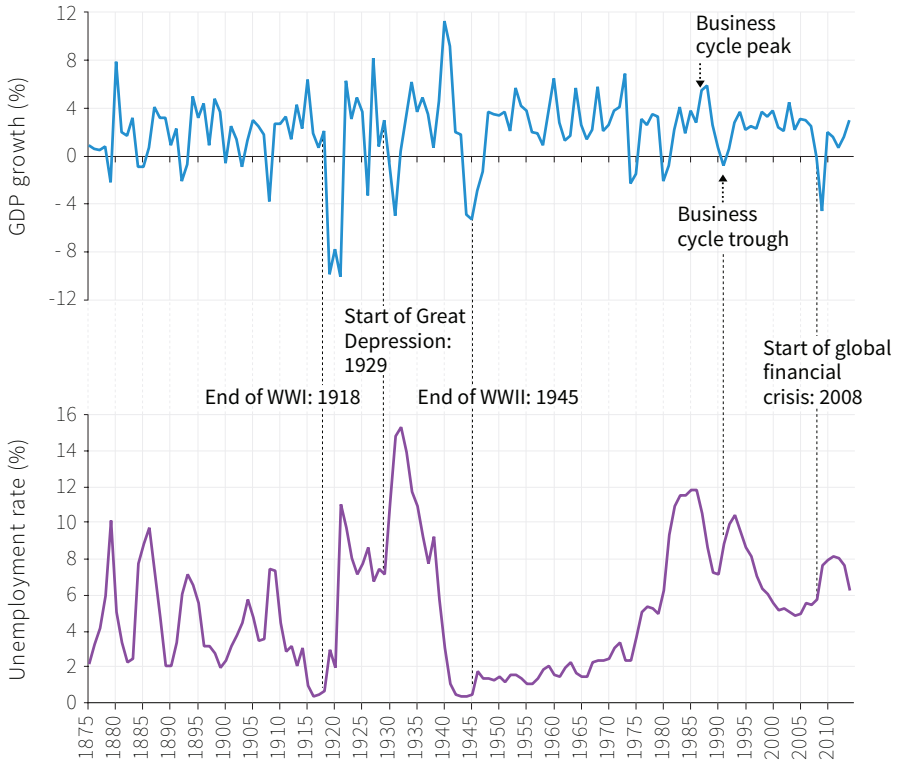
In the right-hand panel, the slope of the line (shown in dashed black) represents the average annual growth rate from 1921 to 2014. It was 2.0% per annum. We can see that growth was steady. A line drawn through the log series from 1875 to 1914 is flatter than the line from 1921: the average growth rate in that period was only 0.9% per annum.

Figure 12.2 *UK GDP per capita (1875-2014).*

Source: *Bank of England. 2015. 'Three Centuries of Macroeconomic Data.'*

We come back to long-run growth in Units 15 and 17. In this unit we focus on fluctuations. These are the jagged lines above and below the dotted black line showing the long-run growth rate in Figure 12.2.

The top panel of Figure 12.3 plots the annual growth rate of UK GDP between 1875 and 2014. Since we want to focus on the size of the economy and how it changes from year to year, we will examine total GDP rather than GDP per capita.



We can see that downturns in the business cycle and rising unemployment are associated. Unemployment continued to rise for a time after the growth rate began to rise in the business cycle in the early 1990s.

Figure 12.3 *UK GDP growth and unemployment rate (1875-2014).*

Source: *Bank of England. 2015. 'Three Centuries of Macroeconomic Data.'*

It is clear from the ups and downs of the series in Figure 12.3 that economic growth is not a smooth process. We often hear about economies going through a boom or a *recession* as growth swings from positive to negative, but there is no standard definition of these words. The National Bureau of Economic Research (NBER) in the US states that: "During a recession, a significant decline in economic activity spreads across the economy and can last from a few months to more than a year." An alternative definition says that an economy is in recession during a period when the level of output is below its normal level. So we have two definitions of recession:

- *NBER definition*: output is declining. A recession is over once the economy begins to grow again.
- *Alternative definition*: the level of output is below its normal level, even if the economy is growing. A recession is not over until output has grown enough to get back to normal.

There is a practical problem with the second definition: it is a matter of judgement, and sometimes controversy, what an economy's normal output would be. (We return to this issue later, where will see that "normal output" is often defined as that consistent with stable inflation.)

DISCUSS 12.2: DEFINING RECESSIONS

A recession can be defined as a period when output is declining, or as a period when the level of output is below normal, sometimes referred to as its *potential* level. Look at this article, especially Figures 5, 6 and 7, to find out more.

1. Consider a country that has been producing a lot of oil and suppose that from one year to the next its oil wells run out. The country will be poorer than previously. According to the two definitions above, is it in a recession?
2. Does knowing whether a country is in recession make a difference to policymakers whose job it is to manage the economy?

The movement from boom, to recession, and back to boom is known as the *business cycle*. The arrows in Figure 12.3 highlight the peak and trough of a business cycle during the late 1980s and early 1990s. You will notice that, in addition to the yearly change in GDP, in which recessions measured by negative growth seem to happen about twice every 10 years, there are less frequent episodes of much larger fluctuations in output. In the 20th century the big downward spikes coincided with the end of the first and second world wars, and with the economic crisis of the Great Depression. In the 21st century, the global financial crisis followed a period in which fluctuations were limited.

In the lower part of Figure 12.3 you can see the unemployment rate. During the Great Depression unemployment in the UK was higher than it had ever been, and it was particularly low during the world wars. Unemployment varies over the business cycle. We can see that downturns in the business cycle and rising unemployment are associated. In the business cycle in the early 1990s, unemployment continued to rise for a time after the economy began to grow again.

12.2 MEASURING THE ECONOMY: UNEMPLOYMENT

According to the standardised definition of the International Labour Organisation (ILO), the *unemployed* are the people who:

- Were without work during the last four weeks, which means they were not in paid employment or self-employment
- Were available for work
- Were seeking work, which means they had taken specific steps in a specified recent period to seek paid employment or self-employment

Figure 12.4 provides an overview of the labour market and shows how the components fit together. We begin on the left-hand side with the population. The next box shows the population of working age. This is total population, minus children and those over 64. It is divided into two parts: the *labour force*, and those out of the labour force (known as inactive). People out of the labour force are not employed or actively looking for work, for example, people unable to work due to sickness or disability, or parents who stay at home to raise children. Members of the labour force can, therefore, be employed or unemployed.

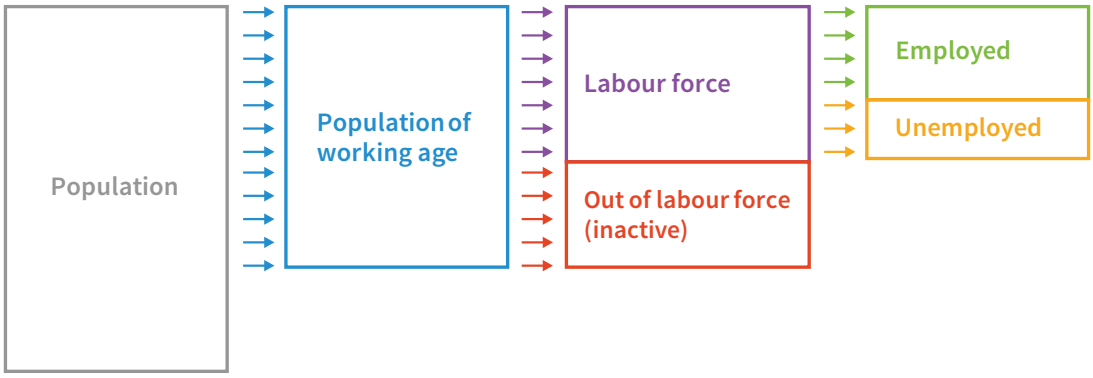


Figure 12.4 *The labour market.*

There are a number of statistics that are useful for evaluating labour market performance in a country, and for comparing labour markets across countries. The statistics depend on the relative sizes of the boxes shown in Figure 12.4. The first is the participation rate, which shows the fraction of the working age population that is in the labour force. It is calculated as follows:

$$\textit{participation rate} \equiv \frac{\textit{labour force}}{\textit{population of working age}}$$

Given that:

$$\textit{labour force} \equiv \textit{employed} + \textit{unemployed}$$

Next is the most commonly cited labour market statistic: the *unemployment rate*. This shows the fraction of the labour force that is unemployed. It is calculated as follows:

$$unemployment\ rate \equiv \frac{unemployed}{labour\ force}$$

Lastly, we come to the *employment rate*, which shows the fraction of the population of working age that are in paid work or self employed. This measure gives us an idea of how well the labour market works within a country. It is calculated as follows:

$$employment\ rate \equiv \frac{employed}{population\ of\ working\ age}$$

It is important to note that the denominator (the statistic on the bottom of the fraction) is *different for the unemployment and the employment rate*. Hence, two countries with the same unemployment rate can differ in their employment rates if one has a high participation rate and the other has a low one.

Figure 12.5 provides a picture of the Norwegian and Spanish labour markets in the 2000s. The figure shows how the labour market statistics relate to each other. It also shows that the structure of the labour market differs widely across countries. We can see that the Norwegian labour market worked better than the Spanish labour market in the 2000s; Norway had a much higher employment rate and a much lower unemployment rate. Norway also had a higher participation rate, which is a reflection of the higher fraction of women in the labour force.

	Norway	Spain
Number of persons, millions		
Population of working age	3.3	36.6
Labour force	2.4	20.6
Out of labour force (inactive)	0.9	16
Employed	2.3	18.3
Unemployed	0.1	2.3
Rates (%)		
Participation rate	2.4/3.3 = 73%	20.6/36.6 = 56%
Employment rate	2.3/3.3 = 70%	18.3/36.6 = 50%
Unemployment rate	0.1/2.4 = 4%	2.3/20.6 = 11%

Figure 12.5 *Labour market statistics for Norway and Spain (2000 to 2009 averages).*

Source: *International Labour Association. 2015. 'ILOSTAT Database.'*

Norway and Spain are illustrations of two common cases. Norway is a low unemployment, high employment economy (the other Scandinavian countries—Sweden and Denmark—are similar) and Spain is a high unemployment, low employment economy (the other southern European economies—Portugal, Italy and Greece—are other examples). Other combinations are possible, however: South Korea is an example of an economy that has both a low unemployment rate and a low employment rate.

DISCUSS 12.3: EMPLOYMENT, UNEMPLOYMENT AND PARTICIPATION

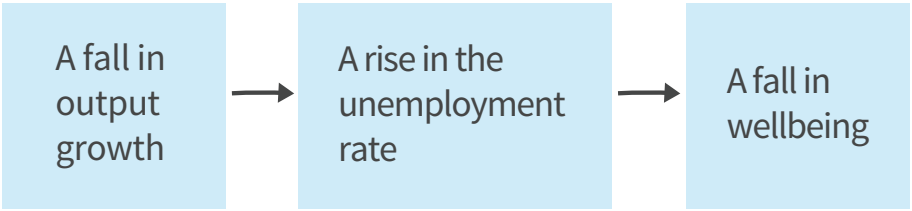
1. Visit the ILO's web site and use the ILOSTAT Database to calculate the employment, unemployment and participation rates for two economies of your choice.
2. Describe the differences in the data for your countries.
3. Can you suggest possible reasons why these differences exist? You may find it useful to find out more about the labour markets.

12.3 OUTPUT GROWTH AND CHANGES IN UNEMPLOYMENT

We saw in Figure 12.3 that unemployment goes down in booms and up in recessions.

Figure 12.6 shows the relationship between output and unemployment fluctuations, known as *Okun's law*. Arthur Okun, an advisor to US President Kennedy, noticed that when a country's output growth was high, unemployment tended to decrease. This is true: Okun's law has been a strong and stable relationship in most economies since the second world war.

We can summarise the relationships like this:



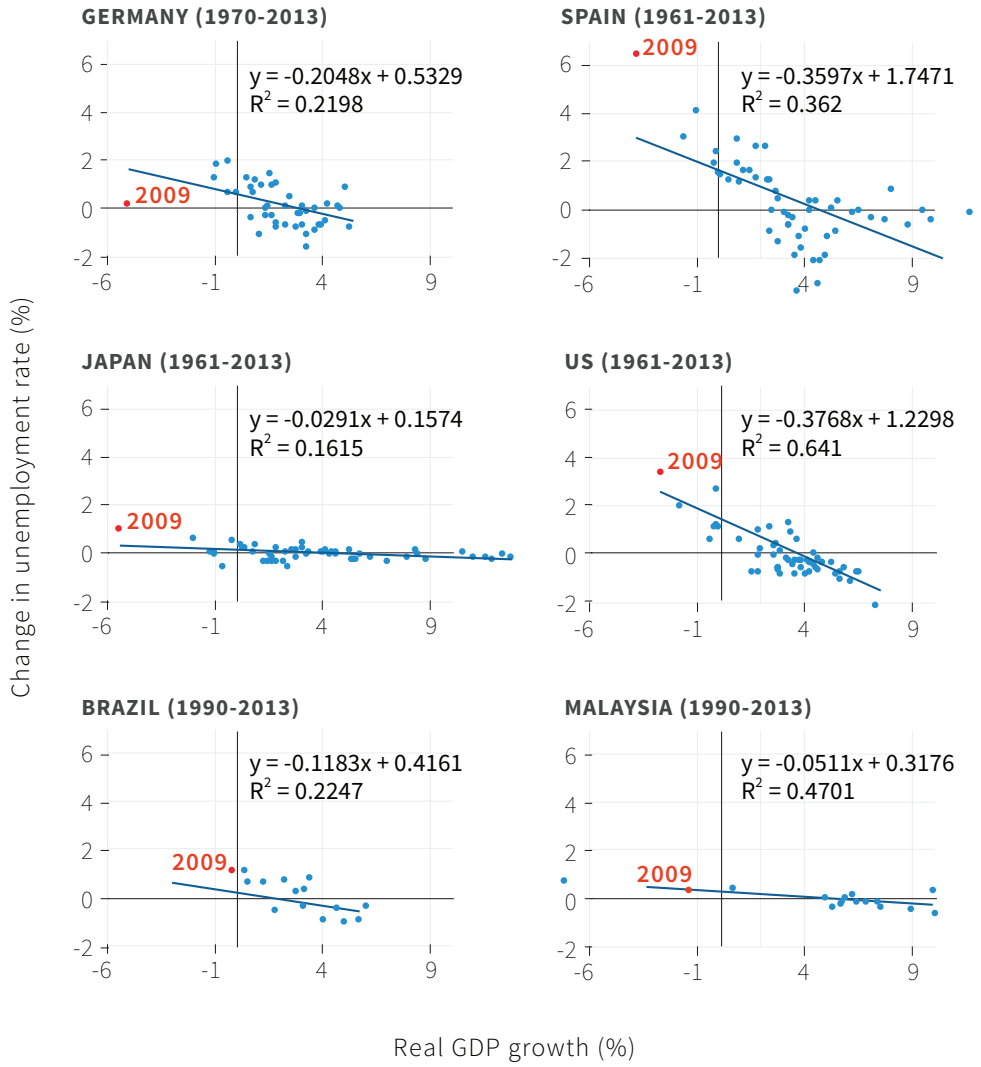


Figure 12.6 *Okun's law, selected economies.*

Source: *OECD. 2015. 'OECD Statistics.'; The World Bank. 2015. 'World Development Indicators.'*

Figure 12.6 plots the change in the unemployment rate on the vertical axis, and the growth rate of output on the horizontal axis for six countries: higher output growth is clearly associated with a decrease in unemployment. In each country chart, there is a downward-sloping line that best fits the points. In the US, for example, the slope of the line implies that, on average, an increase in the output growth rate of 1% decreases the unemployment rate by roughly 0.39%. We say that Okun's coefficient is -0.39 in the US. Our Einstein section shows how to derive the coefficient.

The red circle in each graph in Figure 12.6 shows the changes in real GDP and unemployment that occurred from 2008 to 2009, during the recession that followed the global financial crisis. We can see that, in 2009, all four of the advanced economies experienced their worst output contraction for 50 years. As predicted by Okun's law, unemployment rose substantially by historical standards in Spain, Japan and the US.

But, in each of these three countries, the increase in unemployment was higher than Okun's law predicted. Germany looks very different: Okun's law predicted a rise in unemployment of 1.65 percentage points in Germany but, as the red circle shows, German unemployment hardly changed in 2009. An economic policymaker would surely want to know how Germany managed to protect jobs in the face of the largest decline in the economy's output in 50 years. We will see why this occurred later in this unit.

Brazil and Malaysia also experienced contractions in output and increases in unemployment in 2009. However, like most developing economies, they were hit less hard by the crisis than the advanced economies. Also, Malaysia had recently experienced a much worse contraction during the East Asian crisis in 1998, when growth was -7.4%—bad enough that it would not fit on our chart.

DISCUSS 12.4: OKUN'S LAW

Look again at Figure 12.6—the regression lines for Okun's law all suggest that if growth is equal to zero then unemployment is rising.

1. In these cases, what can you say about labour productivity in the economy?
2. Why does a stable unemployment rate typically require positive growth?

12.4 MEASURING THE AGGREGATE ECONOMY

Economists use what are called *aggregate statistics* to describe the economy as a whole (known as the *aggregate economy*, meaning simply the sum of its parts brought together).

In Figures 12.4 and 12.6 *aggregate output*, or GDP, is the output of all producers in a country, not just those of some region, or in some firm or sector. Recall from Unit 1 that Diane Coyle, an economist who specialises in how we measure GDP, describes it as:

"Everything from nails to toothbrushes, tractors, shoes, haircuts, management consultancy, street cleaning, yoga teaching, plates, bandages, books, and the millions of other services and products in the economy."

The *national accounts* are statistics published by national statistical offices that use information about individual decision-making to construct a quantitative picture of the economy as a whole. There are three different ways to estimate GDP:

- *Spending*: The total spent by households, firms, government and residents of other countries on the home economy's products.
- *Production*: The total produced by the industries that operate in the home economy. Production is measured by the value added by each industry: this means that the cost of goods and services used as inputs to production is subtracted from the value of output. These inputs will be measured in the value added of other industries, which prevents double-counting when measuring production in the economy as a whole.
- *Income*: The sum of all the incomes received as wages, profits and the incomes of the self-employed.

The relationship between spending, production and incomes in the economy as a whole can be represented as a circular flow: the national accounts measurement of GDP can be taken at the spending stage, the production stage, or the income stage. At whichever point the measurement is taken, if accurate measurement were possible, the total of expenditure, output and incomes in a year would be the same.

This is because any *spending* on a good or a service is *income* for whoever sold that output, which also must have been *produced*. If you buy a taco from a street vendor for 20 pesos then your expenditure is 20 pesos, the value added of the taco maker whose production was necessary for this sale is 20 pesos, and the income received by the street vendor is the same 20 pesos. The same point applies if you purchase a car for $20,000, a massage for $50, or insurance for $20 per month.

In 18th century France, a group of economists called *The Physiocrats* studied the economy and compared the way it functioned to the circular flow of blood in the human body. In the economy's circular flow the money flows from the spender to the producer, from the producer to their employees or shareholders, and then is spent again on further output, continuing the cycle, using monetary units (flows of dollars, for example) to construct GDP.

Households and firms both receive income and both spend it. Figure 12.7 shows the circular flow between households and firms, ignoring for now the role of government, or imports and exports.

In the model of the economy in Figure 1.18, we abstracted from the circular flow of income and concentrated attention instead on the physical flows among households, firms and the biosphere. In Unit 18, we look at how the interaction of households and firms with the biosphere can be measured.

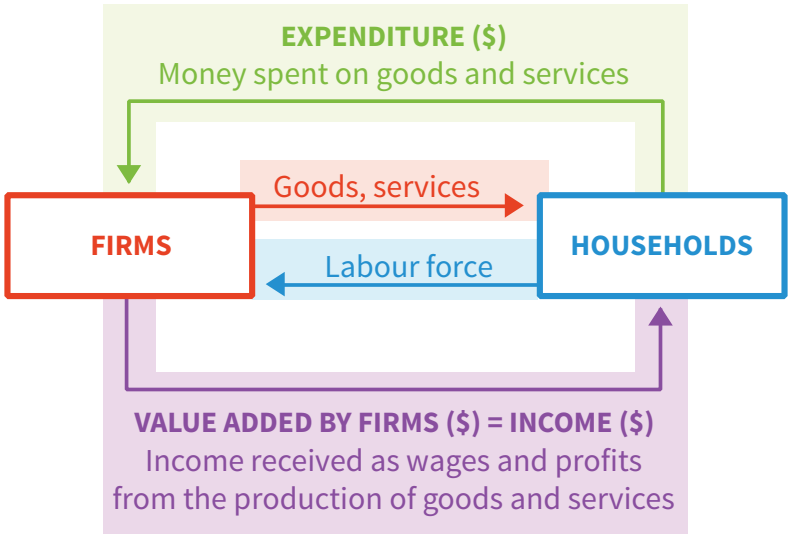


Figure 12.7 *The circular flow model: Three ways to measure GDP.*

GDP can be defined according to any of these three perspectives. But we have to be careful in the definition because, while it is always the case that one person's expenditure is another person's income, globalisation means that often the two people are in different countries. This is the case with imports and exports: someone in China may buy rice from someone in Japan, implying that the expenditure is Chinese while the income is Japanese.

How do we account for these transactions? Since GDP is *domestic* product, it counts as Japanese GDP because the rice was produced (and sold) by Japan. So exports are included in GDP because they are part of domestic production, but imports are not because they are produced elsewhere. For this reason, GDP is defined to *include exports* and *exclude imports*:

- As the value added of domestic production, or as expenditure on domestic production
- As income due to domestic production

The circular flow model shown above considered only households and firms, but the government, and the public services the government provides, can be incorporated in a similar way. Households receive some goods and services that are supplied by the government for which they do not pay at the point of consumption. A good example is primary school education.

The consumption and production of these services can be visualised using the circular flow model:

- *Households to government:* We pay taxes.
- *Government to households:* These taxes pay for the production of public services.

In this way the government can be seen as another producer, like a firm—with the difference that the taxes paid by a particular household pay for public services in general, and do not necessarily correspond to the services received by that household. Since public services are not sold in the market, we also have to make a further assumption: that the value added of government production is equal to the amount it costs the government to produce.

Then we can say that if citizens on average pay $15,000 per year in taxes (the expenditure), that is $15,000 of revenues to the government (the income), which is used to produce $15,000-worth of public goods and services (the value added).

The fact that expenditure, output and incomes are all equal means that we can use one of these perspectives to help us to understand the others. We described recessions as periods of negative output growth. But this means *they must also be periods of negative expenditure growth*—output will only decline if people are buying less. In most recessions, we can even say that output declines *because* people are buying less. This is very useful because we know a lot about what determines expenditure, which in turn helps us to understand recessions, as we will see in Unit 13.

12.5 MEASURING THE AGGREGATE ECONOMY: THE COMPONENTS OF GDP

Figure 12.8 shows the different components of GDP from the spending side as measured in the national accounts for economies on three different continents: the US, the eurozone and China.

Consumption (C)

This includes the goods and services purchased by households. Goods are normally tangible things, in other words, they are things that can be touched. Goods like cars, household appliances and furniture that last for three years or more are called durable goods; those that last less long are non-durable goods. Services are things that households buy that are normally intangible such as transportation, housing (payment of rent), gym membership, and medical services. Household spending on durable goods like cars and household equipment is counted in consumption in the national accounts although, as we will see, in economic terms the decision to buy these long-lasting items is more like an investment decision.

	US	EUROZONE (19 COUNTRIES)	CHINA
CONSUMPTION (C)	68.4%	55.9%	37.3%
GOVERNMENT SPENDING (G)	15.1%	21.1%	14.1%
INVESTMENT (I)	19.1%	19.5%	47.3%
CHANGE IN INVENTORIES	0.4%	0.0%	2.0%
EXPORTS (X)	13.6%	43.9%	26.2%
IMPORTS (M)	16.6%	40.5%	23.8%

Figure 12.8 *Decomposition of GDP in 2013 for the US, eurozone and China.*

Source: *OECD. 2015. 'OECD Statistics.'; The World Bank. 2015. 'World Development Indicators.' OECD reports a statistical discrepancy for China equal to -3.1% of GDP.*

From Figure 12.8 we see that in the advanced countries, consumption is by far the largest component of GDP, close to 56% in the eurozone and 68% in the US. This contrasts with China, where final consumption of households accounts for 37% of GDP.

Investment (I)

This is the spending by firms on new equipment and new commercial buildings; and spending on residential structures (the construction of new housing).

Investment in the unsold output that firms produce is the other part of investment that is recorded as a separate item in the national accounts. It is called the change in inventories or stocks. Including changes in stocks is essential to ensuring that when we measure GDP by the output method (what is produced), it is equal to GDP measured by the expenditure method (what is spent, including investment by firms in unsold inventories).

Investment represents a much lower share of GDP in OECD countries, roughly one-fifth of GDP in the US and the eurozone. In contrast, investment accounts for almost half of GDP in China, the largest share.

Government spending on goods and services (G)

This represents the consumption and investment purchases by the government (consisting of central and local government and called "general government"). Government consumption purchases are of goods (such as office equipment, software, and cars) and services (such as wages of civil servants, armed services, police, teachers, scientists). Government investment spending is on the building of

roads, schools, and defence equipment. Much of government spending on goods and services is for health and education. Government *transfers* in the form of benefits and pensions, such as Medicare in the US, or social security benefits in Europe, are not included in G because households receive them as income: when they are spent, they are recorded in C or I. It would be double-counting to record this spending in G too.

The share of government spending on goods and services is slightly higher in Europe (21.1%) than in the US (15.1%). Remember, this excludes transfers (such as benefits and pensions). The greater difference in the role of the government between Europe and the US comes from those transfers. In 2012 total government spending including transfers was 57% of GDP in France, compared to 36% of GDP in the US.

Exports (X)

Goods and services purchased by households, firms and governments in other countries.

Imports (M)

Goods and services purchased by households, firms and governments in the home economy that are produced in other countries.

AGGREGATE DEMAND

We add up the components of spending in the economy to get GDP: $Y = C + I + G + X - M$. This total is also known as aggregate demand, the total amount of demand for (or expenditure on) goods and services in the economy.

In Figure 12.8 the sum of C, I and G for each of the US, eurozone and China gives the share of GDP accounted for by purchases of goods and services by those living in each of these places. But, taking China as an example, this is not equivalent to the purchases of products made in China. To find that out, first we need to include exports (X): the Chinese goods and services that are purchased by foreigners. Second, since C, I and G include expenditures on imported goods and services, we must subtract total imports (M): the foreign goods and services that are purchased by people living in China.

Net exports

Also called the *trade balance*, this is the difference between the values of exports and imports (X – M).

In 2010, the US had a trade deficit of 3.5% of GDP and China had a trade surplus of 3.8% of GDP. The trade balance is a *deficit* if the number given by exports minus imports as a percentage of GDP is negative; it is called a *trade surplus* if it is positive.

Working with national accounts data is a way of learning about the economy and an easy way to do this is to use FRED (Federal Reserve Economic Data). To learn more about the country where you live and how it compares to other countries, try Discuss 12.5 for yourself.

DISCUSS 12.5: HOW TO USE FRED

If you want real-time macroeconomic data on the German unemployment rate or China's output growth, you do not need to learn German and Chinese, or struggle to get to grips with national archives, because FRED does it for you! FRED is a comprehensive up-to-date data source maintained by the Federal Reserve Board of St Louis in the US, which is part of the US central banking system. It contains the main macroeconomic statistics for almost all developed countries going back to the 1960s. FRED also allows you to create your own graphs and export data into Microsoft Excel.

To learn how to use FRED to find macroeconomic data, use the following steps:

- Click on this link to go to the FRED website
- Use the search bar and type "Gross Domestic Product" (GDP) and the name of a major global economy. Select the annual series for both nominal (current prices) and real (constant prices) GDP for this country. Click the "Add to Graph" button at the bottom of the page. Use the graph that is created to carry out the following tasks:

1. What is the level of nominal GDP in your chosen country this year?
2. FRED tells you that the real GDP is *chained* in a specific year (this means that it is evaluated at constant prices for that year). Note that the real GDP and the nominal GDP series cross at one point. Why does this happen?

From the FRED graph, keep only the real GDP series. FRED shows recessions in shaded areas for the US economy using the NBER definition but not for other economies. For other economies, assume that a recession is defined by two consecutive quarters of negative growth. At the bottom of the graph page, select "Create your own data transformation" and click on "Percent change from one year" (FRED gives you a hint about how to calculate a growth rate at the bottom of the page: notes on growth rate calculation and recessions). The series now shows the percent change in real GDP.

3. How many recessions has this economy undergone?
4. What are the two biggest recessions in terms of length and magnitude?

Now add to the graph the quarterly unemployment rate for your chosen economy (click on "Add data series" under the graph and search for "Unemployment" and your chosen country name).

5. How does the unemployment rate react during the two main recessions you have identified?
6. What was the level of the unemployment rate during the first and the last quarter of negative growth for those two recessions?
7. What do you conclude about the link between recession and the variation in unemployment?

Note: To make sure you understand how these FRED graphs are created, you may want to extract the data in Microsoft Excel and reproduce the growth rate of real GDP, and the evolution of the unemployment rate, on the same graph since 1948.

In most countries private consumption spending makes up the largest share of GDP (see Figure 12.8 to check). Investment spending accounts for a much smaller share (China's very high levels of investment, shown in Figure 12.8, are exceptional). We use the data in the national accounts to calculate how much each component of expenditure contributes towards GDP fluctuations.

Figure 12.9 shows the contributions of the components of expenditure to US GDP growth. The data is for 2009, in the middle of the recession caused by the global financial crisis. We can see that:

- Although investment makes up less than one-fifth of US GDP, it was much more important in accounting for the contraction in the economy than the fall in consumption spending.
- Investment accounted for more than three times the effect on GDP as did consumption, although the latter makes up about 70% of US GDP.
- In contrast to consumption and investment, government expenditure contributed positively to GDP growth; the US government used fiscal stimulus to prop up the economy whilst private sector demand was depressed.
- Net exports also contributed positively to GDP, which reflects both the stronger performance of emerging economies in the aftermath of the crisis and the collapse in import demand that accompanied the recession.

The equation below shows how GDP growth can be broken down into the contributions made by each component of expenditure. We can see that the contribution of each component to GDP growth depends on the share of GDP the component makes up and its growth over the previous period.

$$\text{Percentage change in GDP} = \begin{array}{l} (\text{percentage change in consumption} \times \text{share of consumption in GDP}) \\ + \\ (\text{percentage change in investment} \times \text{share of investment in GDP}) \\ + \\ (\text{percentage change in government spending} \times \text{share of government spending in GDP}) \\ + \\ (\text{percentage change in net exports} \times \text{share of net exports in GDP}) \end{array}$$

	GDP	CONSUMPTION	INVESTMENT	GOVERNMENT SPENDING	NET EXPORTS
2009	-2.8	-1.06	-3.52	0.64	1.14

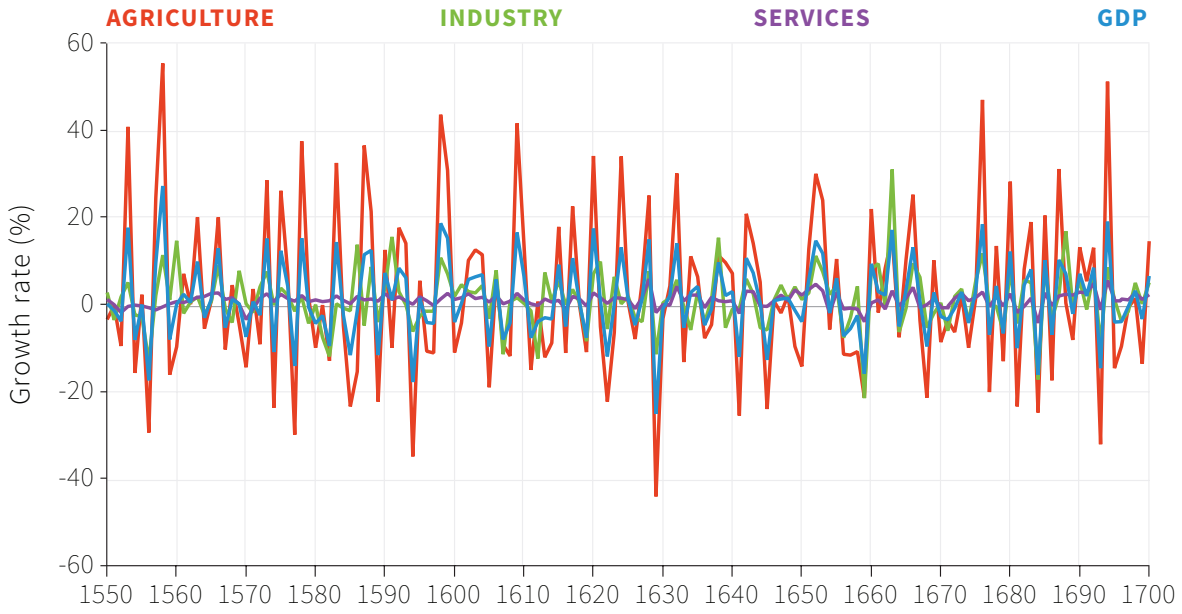
Figure 12.9 *Contributions to percentage change in real GDP in the US in 2009.*

Source: *Federal Reserve Bank of St. Louis. 2015. 'FRED.' Note that government investment comes under government spending, and not investment, in the national accounts.*

12.6 HOUSEHOLDS COPE WITH FLUCTUATIONS

Economies fluctuate between good and bad times; so far we have studied industrialised economies, but this is equally true in economies based on agriculture. Figure 12.10a illustrates fluctuations in the largely agrarian British economy between 1550 and 1700. It shows the growth rate of real GDP and of the three main sectors: agriculture, industry and services. Clearly the agricultural sector is much more volatile than the other sectors of the economy. Since agriculture also accounted for the highest share of GDP during this period, the agricultural sector largely drove fluctuations in GDP.

Figure 12.10b shows the growth rates of real GDP and agriculture in India since 1960. In 1960 agriculture comprised 43% of the economy, which had declined to 17% in 2014. Partly due to modern farming methods, agriculture in modern India is not as volatile as it was in Britain before 1700. But it remains nearly twice as volatile as GDP as a whole.



In this period the average difference in the output of the agricultural sector from one year to the next is three times larger than that of the industrial sector and more than 10 times larger than that of the services sector.

Figure 12.10a *The role of agriculture in the fluctuations of the aggregate economy in Britain (1550-1700).*

Source: *Broadberry, Stephen, Bruce M. S. Campbell, and Alexander Klein. 2015. British Economic Growth, 1270-1870. Cambridge: Cambridge University Press.*

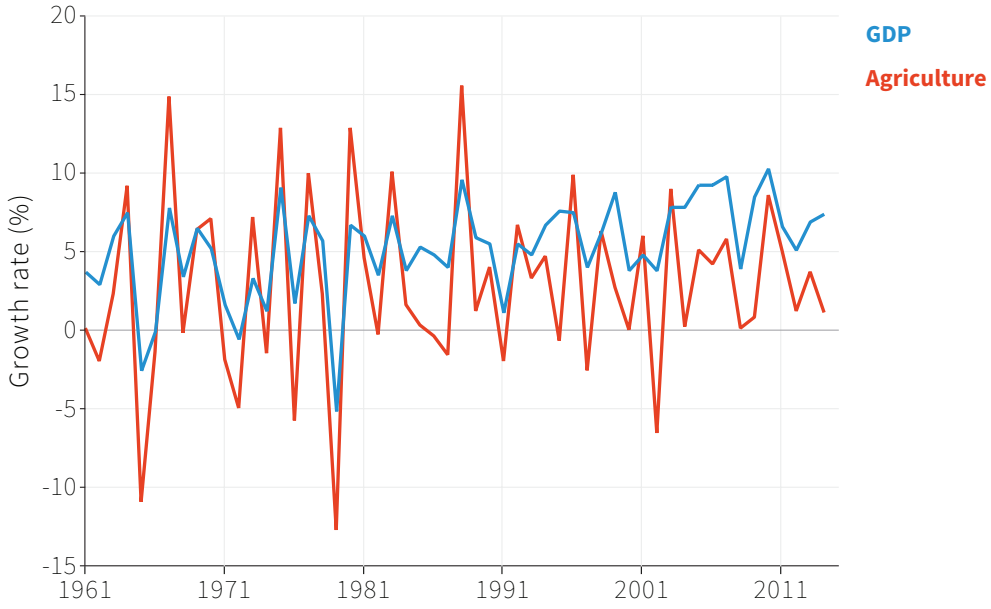


Figure 12.10b *The role of agriculture in the fluctuations of the aggregate economy in India (1961-2014).*

Source: *The World Bank. 2015. 'World Development Indicators.'*

To help us to think about the costs and causes of economic fluctuations, we begin with an agrarian economy. In an economy based on agricultural production, the weather—with war and disease—is a major cause of good and bad years. The term *shock* is used in economics to refer to an unexpected event—for example extreme weather, or a war. As we know, people think about the future and usually they anticipate that unpredictable events may occur. They also act on these beliefs. In a modern economy, this is the basis of the insurance industry. In an agrarian economy, households also anticipate that both bad luck and good harvests can occur.

How do households cope with fluctuations that can cut their income in half from one season to the next? We can distinguish between two situations:

- *Good or bad fortune strikes the household:* For example when disease affects a family's animals, or when a family member who plays an important role in farming is injured.
- *Good or bad fortune strikes the economy as a whole:* For example when drought, disease, floods, a war or an earthquake affects a whole area.

Household shocks

People use two strategies to deal with shocks that are specific to their household:

- *Self-insurance:* Households that encounter an unusually high income in some period will save; when their luck reverses, they spend their savings. They may also borrow in bad times if they can, depending on how credit-constrained they are—as we saw in Unit 11. It is called self-insurance because other households are not involved.
- *Co-insurance:* Households that have been fortunate during a particular period help a household hit by bad luck. To some extent this is done among members of extended families and among friends and neighbours. During the last half century, particularly in richer countries, co-insurance has taken the form of citizens paying taxes which are then used to support individuals who are temporarily out of work, called unemployment benefits or insurance.

Informal co-insurance among family and friends is based on both reciprocity and trust: you are willing to help those who have helped you in the past, and you trust the people who you helped to do the same in return. *Altruism* towards those in need is also usually involved, although co-insurance can work without it.

These strategies reflect two important aspects of household preferences:

- *People prefer a smooth pattern of consumption:* They dislike consumption that fluctuates as a result of bad or good shocks such as injury or good harvests. So they will self-insure.
- *Households are not solely selfish:* They are willing to provide support to each other to help smooth the effect of good and bad luck. They often trust others to do the same, even when they do not have a way of enforcing this. Altruistic and

reciprocal preferences remain important even when co-insurance takes the form of a tax-supported employment benefit because these are among the motives for supporting the public policies in question.

Economy-wide shocks

But co-insurance does not work if the bad shock hits everyone at the same time. When there is a drought, flood or earthquake, there are few ways an agrarian economy can protect the wellbeing of the people who are affected. For example, it is not usually possible to store produce from a bumper harvest for several years, until a bad harvest occurs.

In farming economies of the past that were based in volatile climates, people practiced co-insurance based on trust, reciprocity and altruism. These are norms, like the fairness norm we discussed in Unit 4, and they probably emerged and persisted because they helped people to survive in these regions that were often hit by bad weather shocks. Recent research suggests that they seem to have persisted even after climate had become largely unimportant for economic activity.

The evidence for this is that people in the regions with high year-to-year variability in rainfall and temperature during the past 500 years now display high levels of trust, and have more modern day co-insurance institutions such as unemployment insurance (unemployment benefit payments) and government assistance for the disabled and poor.

DISCUSS 12.6: HEALTH INSURANCE

Think about the health insurance system in your country.

1. Is this an example of co-insurance or self-insurance?
2. Can you think of other examples of both co-insurance and self-insurance?

In each case, consider what kinds of shocks are being insured against and how the scheme is financed.

12.7 WHY IS CONSUMPTION SMOOTH?

A basic source of stabilisation in any economy comes from the desire of households to keep the level of their consumption of goods and services constant. Keeping a steady level of consumption means households have to plan. They think about what might happen to their income in the future, and they save and borrow to smooth the bumps in income.

We have seen that this behaviour occurs in agrarian societies faced by weather and war shocks, but modern households also try to smooth their consumption. One way to visualise this behaviour is to focus on predictable events. A young person thinking about life can imagine getting a job, then enjoying a period of working life with income higher than the starting salary, followed by years in retirement when income is lower than during working life.

As we saw in Unit 11, people prefer to smooth their consumption because there are diminishing marginal returns to consumption at any given time. So it will be the case that having a lot of consumption later and little now, for example, is worse than having some intermediate amount of consumption in the two periods (Figure 11.3a).

The person contemplating a future promotion and planning spending would be in a position similar to Julia in Unit 11 (Figure 11.1) who had limited funds in the present but knew she would have more later, and consequently was interested in moving some of her future buying power to the present by borrowing. The model of decision-making for the individual that we introduced in Unit 3 and Unit 11 is the basis for thinking about consumption throughout a person's life: it predicts that, although income fluctuates throughout our lives, our desired consumption is smoother.

We can use Figure 12.11 to visualise an individual's tendency to smooth consumption expenditure. The blue line shows the path of income over time: it starts low, rises when the individual is promoted and falls at retirement. Consumption expenditure is the red line: this is smooth (flat) from the point at which the individual first gets a job. This is the point at which this individual makes a lifetime consumption plan based on a forecast about how income will change. In this simple example, before starting work, the individual's income and consumption expenditure are the same—we assume, for example, that parents support their children until the children start work.

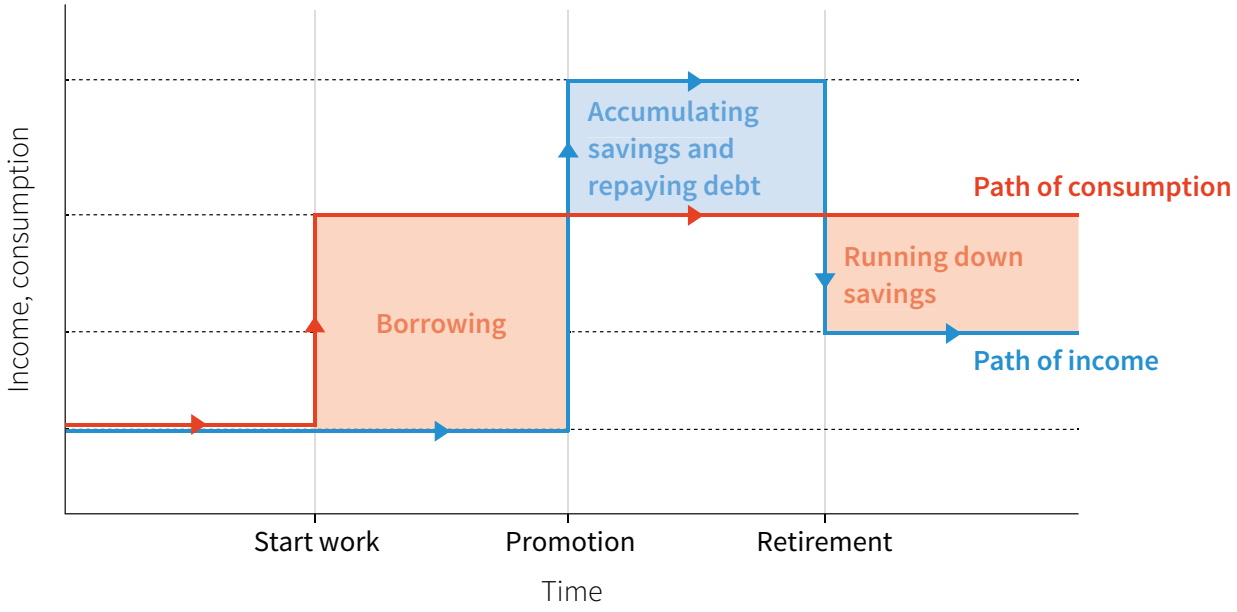


Figure 12.11 *Consumption smoothing through our lifetime.*

The notable feature of Figure 12.11 is that consumption changes before income does. Like a family in an agrarian economy that begins saving for a daughter's dowry before she is old enough to marry, the individual shown in Figure 12.11 thinks ahead to high income after a promotion, and adjusts consumption upward ahead of time. As we have seen in Unit 11, this implies that the individual can borrow. Maybe it is possible to convince the bank that the job is secure and prospects are good. If so, the individual can probably get a mortgage now, and live in a more comfortable house with a higher standard of living than would be the case if long-term earnings were to remain at the starting salary. The labels on Figure 12.11 show that the individual borrows while young and income is low, saves and repays the debt when older and earning more, and finally runs down savings after retirement, when income falls again.

The model of decision-making in Figure 12.11 highlights the desire of households for a smooth path of consumption. We next ask what happens when something *unexpected* occurs to disturb the lifetime consumption plan? What if the individual shown in Figure 12.11 encounters an unexpected income shock? The consumption-smoothing model suggests that:

- First, the individual will make a judgement about whether the shock is temporary or permanent.
- *If the shock is permanent:* We readjust the red line in Figure 12.11 up or down to reflect the new long-run level of consumption that the individual adopts, consistent with the new pattern of forecast income.
- *If the shock is temporary:* Little will change. A temporary fluctuation in income has almost no effect on the lifetime consumption plan, because it makes only a small change to lifetime income.

To summarise, when individuals and households behave in the way shown in Figure 12.11, shocks to the economy will be dampened because spending decisions are based on long-term considerations. They aim to avoid fluctuations in consumption even when income fluctuates. Read how Daryl Collins, an economist, and his collaborators document the way some very poor households manage their finances as they attempt to avoid living from hand-to-mouth.

What limits a household's consumption smoothing? Many individuals and households are not able to make or implement long-term consumption plans. Making plans can be difficult because of a lack of information or, even if we have information, we can't use it to predict the future with confidence. For example, it is often very hard to judge whether a change in circumstances is temporary or permanent.

There are three other things that constrain the ways in which households can smooth their consumption when faced with income shocks. The first two concern limits on self-insurance, the third is a limit on co-insurance:

- *Credit constraints or credit market exclusion:* Introduced in Unit 11, this restricts a family's borrowing to sustain consumption when income has fallen.
- *Weakness of will:* A characteristic of human behaviour that leads people to be unable to carry out the plans—for example saving in anticipation of negative income shock—that they know would make them better off.
- *Limited co-insurance:* So that those with a fall in income cannot expect much support in sustaining their incomes from others more fortunate than them.

Credit constraints

As we saw in Unit 11, the amount a family can borrow is limited, particularly if it is not wealthy. Households with little money cannot borrow at all, or only at extraordinarily high interest rates. Thus the people who most need credit to smooth their consumption often are unable to do so. The credit constraints and credit market exclusion discussed in Unit 11 help explain why borrowing is often not possible.

Figure 12.12 shows the reaction of two different types of households to an anticipated rise in income. Households that are able to borrow as much as they like are in the top panel. Credit-constrained households that are unable to get a loan or take out a credit card are in the bottom panel. We have marked two key events. First, the household receives news that income will rise at a predictable time in the future (the news may be about a promotion or a bequest, for example). Second, the household's income rises (the promotion happens, the inheritance comes through). The blue lines on the figure show that the path of income over time is the same in both households. The red line in the top panel shows that, in a consumption-smoothing household, consumption changes immediately it gets the news. On the other hand, a credit-constrained household that cannot borrow has to wait until the income arrives before adjusting its standard of living.

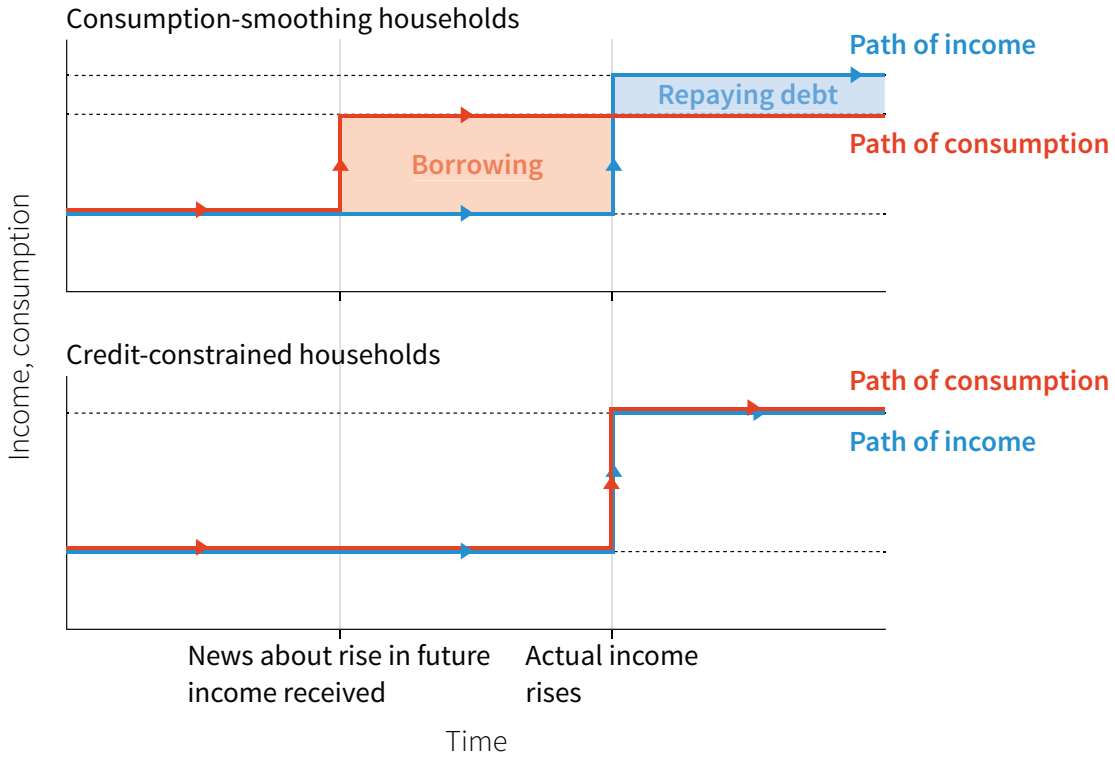


Figure 12.12 *Consumption when credit constraints bind: An anticipated rise in income.*

We can think about these decisions using the two-period model of borrowing and lending from Unit 11, shown in Figure 12.13. First consider a household that receives the same income, y, this period and next period, indicated by the endowment point A in Figure 12.13. The interest rate is r so if the household can borrow and save then it can consume anywhere on the budget constraint, which has the slope $-(1+r)$. The budget constraint is another term for the frontier of the feasible set with the slope of $-(1+r)$ which we used in Unit 11.

To focus on credit constraints, we assume the household prefers to consume the same amount each period, shown by the point where the indifference curve is tangent to the budget constraint (this household is not characterised by pure impatience). So it simply consumes its income y in both periods and does not want to borrow or save: it is at its endowment point, A.

Now suppose that the household experiences an unexpected negative shock to its income this year—such as a bad harvest—which lowers this year's income to y'. However, this is expected to be a temporary shock: it expects income next year to return to the higher level, y. So the household's endowment is now at point A', to the left of point A. Again, if it can borrow and save then its budget constraint has a slope of $-(1+r)$ and passes through point A'. The highest indifference curve that touches this budget constraint does so at point A'' showing that the household prefers to smooth consumption, consuming c' in both periods. The household borrows $c'-y'$ now and repays $(1+r)(c'-y')$ next period.

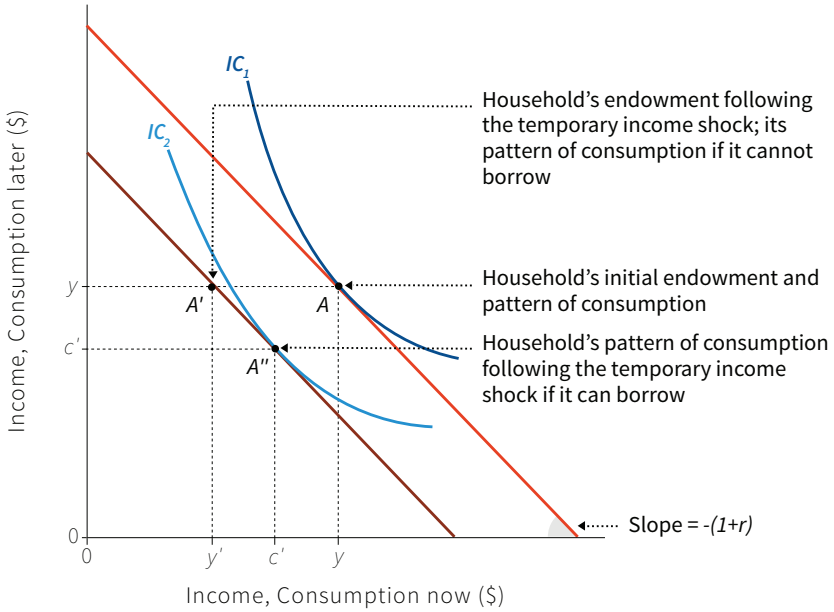


Figure 12.13 *Credit-constrained and unconstrained households: An unanticipated temporary fall in income.*

If the household is credit-constrained, however, then it cannot borrow. It can only consume income y' now. Since it is not borrowing it follows that next period it can consume the whole future income of y, so it consumes at point A'. We learn from this example:

- Without borrowing or lending, the endowment point and pattern of consumption coincide.
- Compared with the smoothing household, the credit-constrained household consumes less this period and more next period.

But we can see that the indifference curve that passes through A' (not shown) is lower than the one that passes through A''. So the household that smoothes consumption by borrowing is better off than the credit-constrained household.

A temporary change in income affects the current consumption of credit-constrained households more than it does that of the unconstrained.

DISCUSS 12.7: CHANGE IN INCOME, CHANGE IN CONSUMPTION

Take the situation in Figure 12.13. Begin at point *A′* for the credit-constrained and at point A for the smoothing households.

1. Explain the relationship between the change in income and the change in consumption when income returns to normal after the temporary decline for each household type.
2. Based on this analysis, explain the predicted relationship between temporary changes in income and consumption for an economy with a mixture of household types.

Weakness of will

In Figure 12.14, an individual learns that income is going to fall in the future. This could be because of retirement or job loss. It could also be because the individual is becoming pessimistic: perhaps the newspapers predict an economic crisis. In the top panel of Figure 12.14 we again show a household behaving in a forward-looking manner to smooth consumption.

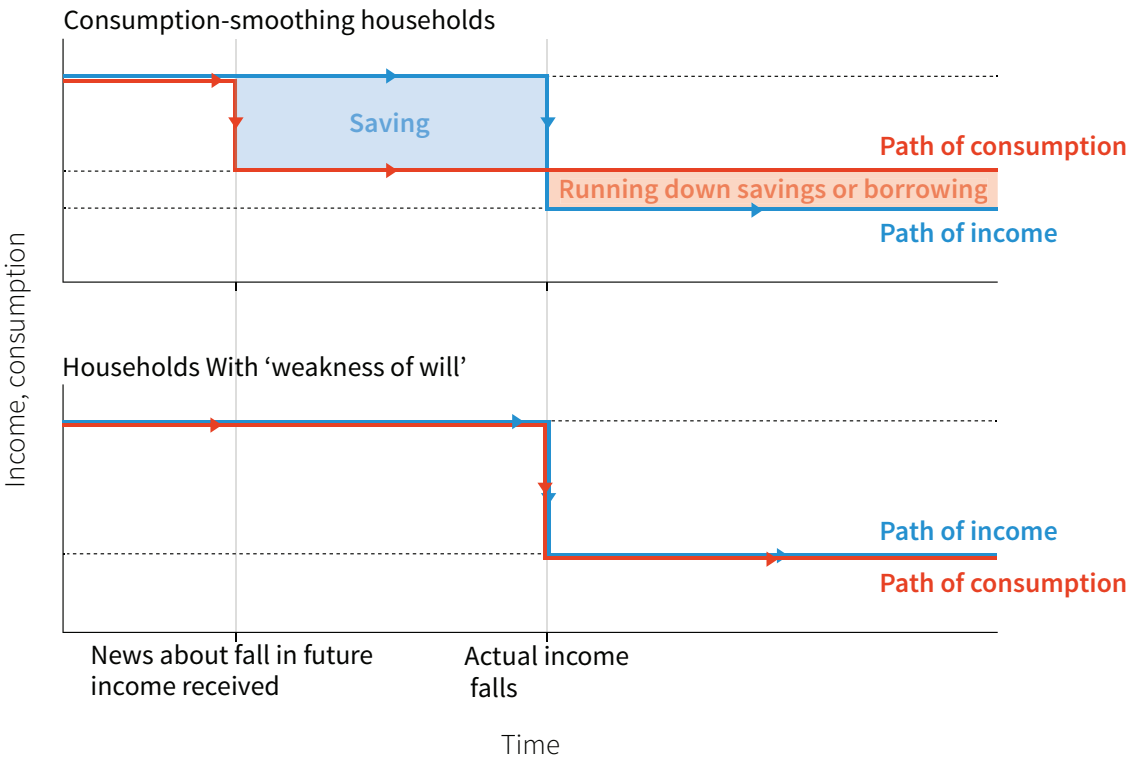


Figure 12.14 *Consumption when households are weak-willed: An anticipated fall in income.*

The bottom panel of Figure 12.14 is a household that consumes too much today, even if it implies a large reduction in consumption in the future. The blue lines in the figure show that income follows the same path in both sets of households. The red line in the top panel shows the consumption path for a consumption-smoothing household. When it receives news of the imminent fall in income, it immediately starts saving to supplement consumption when income falls. In contrast, the weak-willed household does not react to the news, and keeps consumption high until income falls.

WEAKNESS OF WILL

This describes the inability to commit to a course of action (dieting or foregoing some other present pleasure for example) that one will regret later. Impatience may lead a person to favour pleasures in the present, but not necessarily to act in a way that one regrets.

This feature of human behaviour is familiar to many of us: we often lack willpower. The problem of not being able to save obviously differs from the problem of not being able to borrow: saving is a form of self-insurance and doesn't involve anyone else.

MY DIET STARTS TOMORROW

Economists have conducted experiments to test for behaviour that would help to explain why we don't save when we can. For example, Daniel Read and Barbara van Leeuwen conducted an experiment with 200 employees at firms in Amsterdam. They asked them to choose today what they thought they would eat next week. The choice was between fruit and chocolate.

When asked, 75% of subjects replied that they would eat fruit next week. When asked what they would choose to eat today, 70% chose chocolate. But, when next week came, only 30% actually chose to eat fruit. The experiment shows that, although people may plan to do something that they know will be beneficial (eat fruit, save money), when the time comes they often don't do it.

Read, Daniel, and Barbara van Leeuwen. 1998. 'Predicting Hunger: The Effects of Appetite and Delay on Choice.' *Organizational Behavior and Human Decision Processes* 76 (2): 189–205.

Limited co-insurance

Most households lack a network of family and friends who can help out in substantial ways over a long period when a negative income shock occurs. As we have seen, unemployment benefits provide this kind of co-insurance—the citizens who turn out to be lucky in one year insure those who are unlucky. But in many societies the coverage of these policies is very limited.

A vivid demonstration of the value of smoothing through co-insurance is the experience of Germany during the drastic reduction in income experience by that economy in 2009 (see Figure 12.6). When the demand for firms' products fell, workers' hours of work were cut but very few Germans lost their jobs, and many of those at work continued to be paid as if they were working many more hours than they did as a result of both government policy and agreements between firms and their employees. The result was that though income had fallen, consumption did not—and unemployment did not increase.

But most empirical evidence shows that credit constraints, weakness of will and limited co-insurance mean that, for many households, a change in income results in an equal change in consumption. In the case of a negative income shock such as the loss of a job, this means that the income shock will now be passed on to other families who would have produced and sold the consumption goods that are now not demanded.

We will see in the next unit how the initial shock in income may be multiplied (or amplified) by the fact that families are limited in their ability to smooth their consumption. This in turn helps us understand the business cycle and how policymakers may or may not help to manage it.

12.8 WHY IS INVESTMENT VOLATILE?

Households tend to smooth their consumption spending when they can: put simply, because they have to eat. But there is no similar motivation for a firm to smooth investment spending. Firms increase their stock of machinery and equipment and build new premises when they see an opportunity to make profits. But, unlike eating and most other consumption expenditures, investment expenditures can be postponed. There are several reasons why this is likely to produce clusters of investment projects at some times, while at other times there are few.

In Unit 2, we saw how firms responded to profit opportunities in the Industrial Revolution by innovating. This helps explain why investment occurs in waves. When an innovation like the spinning jenny is introduced, firms using the new

technology can produce output at lower cost or produce higher-quality output. They expand their share of the market. Firms that fail to follow may be forced out of business because they are unable to make a profit using the old technology. But new technology means firms must install new machines. As firms do this, there is an investment boom. This will be amplified if the firms producing the machinery and equipment need to expand their own production facilities to meet the forecast extra demand.

In this case investment by one firm pushes other firms to invest: if they don't, they may lose market share or even be unable to cover their costs and eventually have to leave the industry. But investment by one firm can also pull other firms to invest by helping to increase their market and potential profits.

An example of push investment is the hi-tech investment boom in the US: from the mid-1990s, new information and communications technology (ICT) was introduced into the US economy on a large scale. Figure 12.15 shows the sustained growth of investment in new technologies through the second half of the 1990s.

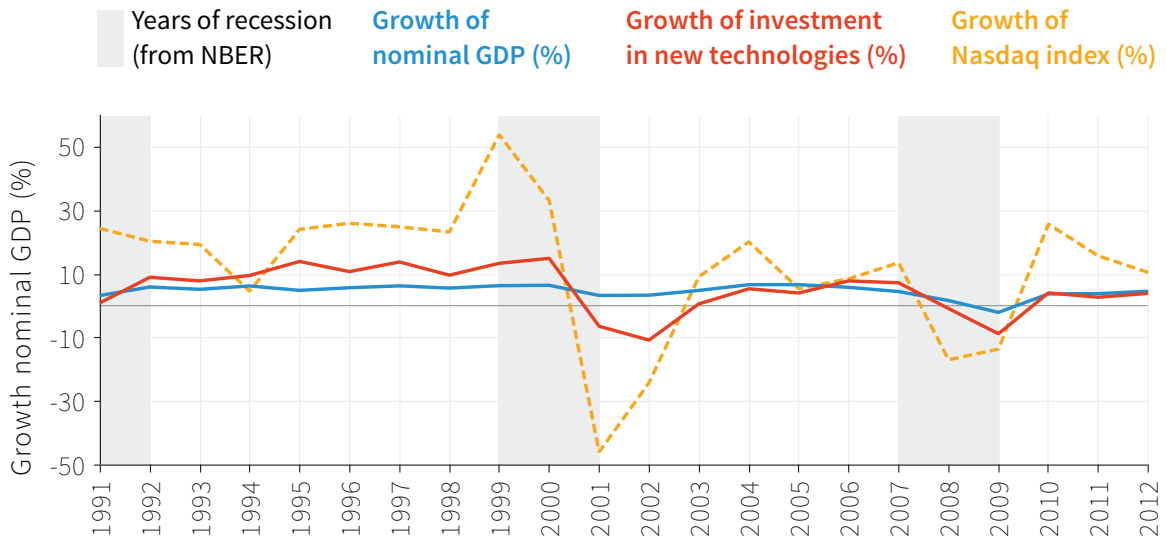


Figure 12.15 *Investment in new technologies and the dotcom bubble (1991-2012).*

Source: *US Bureau of Economic Analysis. 2015. 'Fixed Assets Accounts Tables.'*

Note: *the series are in current US dollars. Nasdaq value is the yearly average of the close price value of the Nasdaq. Investment in new technologies is the investment in information processing equipment, computers and peripheral equipment, communication equipment, communication structure, and IPPR investments for software, semi-conductors and other electronic component and computers.*

As we saw in Unit 9, investment in new technology can lead to a bubble on the stock market and to over-investment in purchases of machinery and equipment. The chart shows in yellow the behaviour of the stock market index in the US on which hi-tech companies are listed. This is the Nasdaq index, introduced in Unit 9.

The index rises strongly from the mid-1990s to an all-time peak in 1999 as stock market investors' confidence in the profitability of new tech firms grew. Investment in IT equipment (the red line) grew rapidly as a result of this confidence, but dropped sharply following the collapse in confidence that caused the fall of the stock market index. This suggests that over-investment in machinery and equipment had occurred: investment did not return to growth until 2003. Economist Robert Shiller argued that the Nasdaq index was driven high by what he called "irrational exuberance", as you might recall from Unit 9. Beliefs in the future of hi-tech led not only to share prices rising to levels that were unsustainable, but also to excessive investment in machinery and equipment in the hi-tech sector.

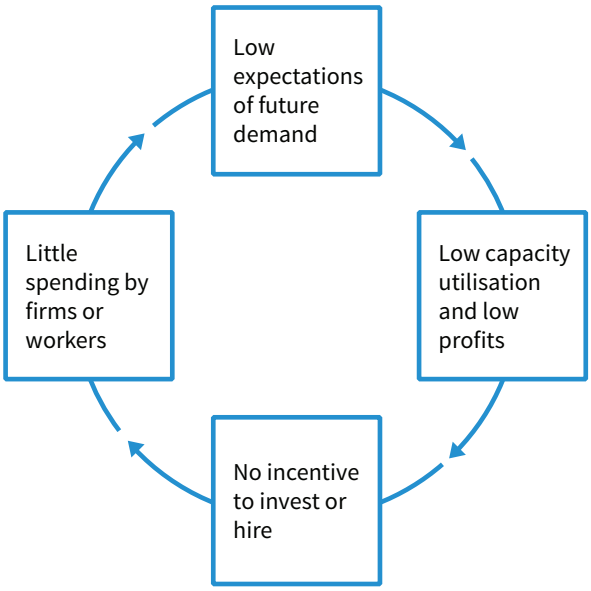
CAPACITY UTILISATION RATE

This is a measure of the extent to which a firm, industry or entire economy is producing as much as the stock of its capital goods and current knowledge would allow. Low capacity utilisation, say 70%, indicates that the economy is currently producing only 70% of what it could do without investing in new buildings and equipment, if the demand for its products were sufficient to justify hiring more labour.

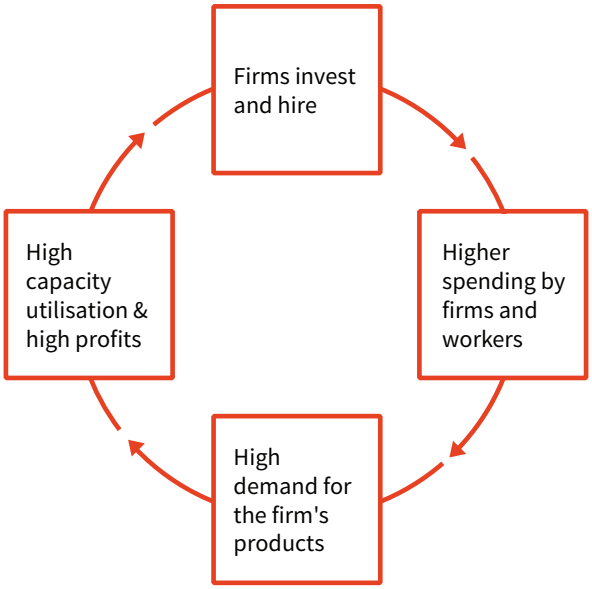
In Unit 11 we saw that the ability of households and firms to implement their plans depends on their ability to borrow, that is, on their access to credit. Credit constraints are therefore another reason for the clustering of investment projects and the volatility of aggregate investment. In a buoyant economy, profits are high and firms can use these profits to finance investment projects. Access to external finance from sources outside the firm is also easier: in the hi-tech boom in the US, for example, the expansion of the Nasdaq exchange reflected the appetite of investors to provide finance by buying shares (stocks) in firms in the emerging ICT industries.

To understand how one firm's investment can pull another to invest, think of a local economy comprising just two firms. Firm *A*'s machinery and equipment are not fully used, so the firm can produce more if it hires more employees. But there is not enough demand to sell the products it would produce. This situation is called low *capacity utilisation*. The owners of Firm *A* have no incentive to hire more workers or to install additional machinery: that is, to invest.

Firm *B* has the same problem. Because of low capacity utilisation, profits are low for both. Thus when we think about both firms together we have a vicious circle:



If the owners of both *A* and *B* decide to invest and hire at the same time, they would employ more workers, who would spend more, increasing the demand for the products of both firms. The profits of both would rise, and we have a virtuous circle:



These two circles highlight the role of expectations of future demand, which depend on the behaviour of other actors. How to get out of the vicious circle and into the virtuous one can be studied using a game similar to those studied in Unit 4. As in every game we specify:

- *The actors:* The two firms.
- *The actions that they can take:* Invest, or do not invest.
- *The information they have:* They decide simultaneously, so they do not know what the other has done.

- *The payoff*: The profits resulting from each of the four pairs of actions that they could possibly take.

The four possible outcomes of the interaction and the payoffs are given in Figure 12.16.

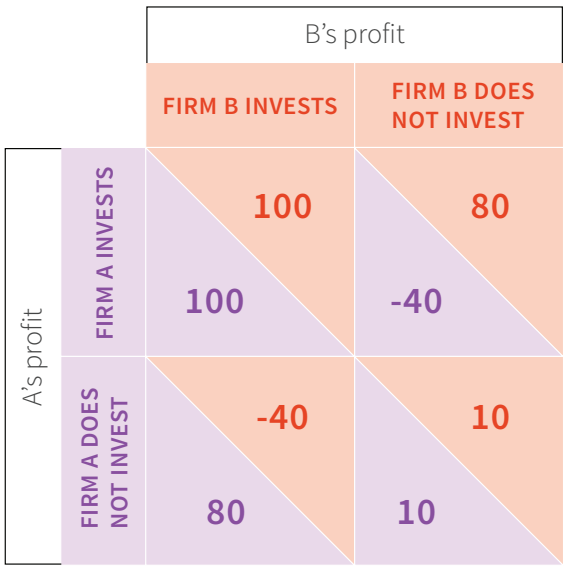


Figure 12.16 *Investment decisions as a coordination game.*

From the figure you can see what happens when the virtuous (both invest) and vicious (both do not invest) circles occur. Note what happens if one of the firms invests but the other does not. If firm *A* invests and *B* does not (the upper right cell in the figure) then *A* pays to install new equipment and premises, but because the other firm did not invest there is no demand for the products that the new capacity could produce; so *A* makes a loss. But had *B* known that *A* would invest, then *B* would have made higher profits by investing as well (getting 100 rather than only 80). On the other hand, had *B* known that *A* was not going to invest, then it would have done better to also not invest.

In this game the two firms will do better if they do the same thing, and best of all if that thing is to invest. This is another reason that investment tends to fluctuate a lot. If owners of firms think that other firms will not invest, then they will not invest—confirming the pessimism of the other owners. This is why the vicious circle is self-reinforcing. The virtuous circle is self-reinforcing for the same reason: optimism about what other firms will do leads to investment, which sustains the optimism.

There are two *Nash equilibria* in this game (upper left and lower right). The Nash equilibrium (lower right) in which both firms have low capacity utilisation, and low hiring and investment, is not Pareto efficient because there is a change in which both make higher profits, namely if both firms decide to invest. This situation is like the driving on the right or left side of the road game, discussed in Unit 4, or the interaction described in Figure 4.14 concerning specialisation in different crops, or global climate change described in Figure 4.17. These are all called *coordination games*.

The name is very apt here because to make the move from the vicious to the virtuous circle, the firms have to coordinate in some way (both agree to invest) or develop optimistic beliefs about what the other will do. This kind of optimism is often called business confidence, and it has a major role in the fluctuations in the economy as a whole. Under some circumstances government policy, as we will see in the next unit, can also help shift an economy from the Pareto-inefficient outcome to the Pareto-efficient outcome.

COORDINATION GAME

A game in which there are two Nash equilibria and in which one may be Pareto superior to the other is called a *coordination game*.

- Driving on the right or the left is a coordination game in which neither equilibrium is preferable to either player.
- In the crop specialisation coordination game in Unit 4 (Figure 4.14) specialisation in the "right" crops (a different crop for the two farmers) is better for both than the "wrong specialisation".
- In the investment coordination game (Figure 12.16), an outcome in which both invest is better for both than neither investing.

We can generalise the argument about the role of coordination in investment to say that investment spending by firms will respond positively to the growth of demand in the economy. Once an increase in aggregate spending on home's production of goods and services (that is, on $C + I + G + X - M$) occurs, this helps to coordinate the forward-looking plans of firms about their future capacity needs, and stimulates investment spending.

Figure 12.17 illustrates the relationship between the growth of aggregate demand (excluding investment), business confidence, and investment for the eurozone. The business confidence indicator moves closely with aggregate demand (excluding investment) and investment.

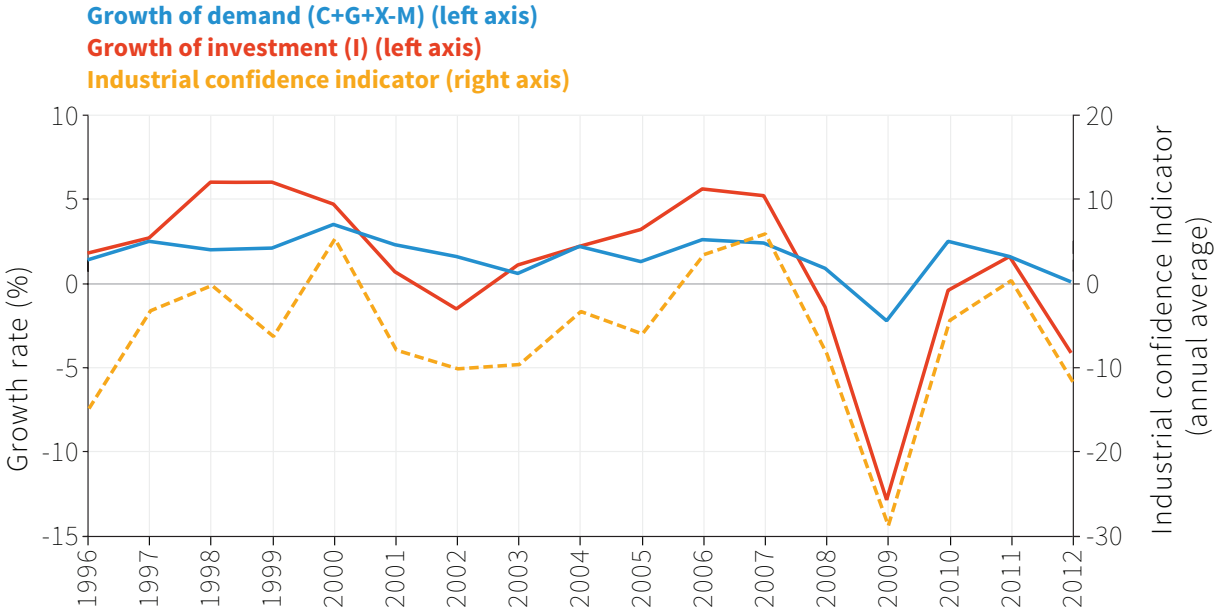


Figure 12.17 *Investment and business confidence in the eurozone (1996-2012).*

Source: *Eurostat. 2015. 'Confidence Indicators by Sector.'; Federal Reserve Bank of St. Louis. 2015. 'FRED.'*

Therefore we would expect the data from the national accounts to confirm that consumption spending is smoother and investment spending more volatile than GDP in the economy as a whole.

As expected, Figures 12.18a and 12.18b show that investment is much more volatile than consumption in two rich countries, the UK and the US, and two middle-income countries, Mexico and South Africa. The upward and downward spikes in the red series for investment are larger than those for the green series for consumption.

A close look at the charts for the rich countries also shows that, as predicted, consumption is less volatile than GDP: the blue peaks and troughs for GDP are larger than the green ones for consumption. This is less evident in the middle-income countries, perhaps because households are more credit-constrained and therefore are less able to borrow in order to smooth their consumption.

How volatile is government spending? Unlike investment, government spending (the G in the national accounts) does not respond to innovation or fluctuate with business confidence. We would predict it to be less volatile than investment. And net exports? The demand for exports will fluctuate with the business cycle in other countries, and will be affected more by the booms and recessions of the countries that are large export markets. Find out about the volatility of government spending and net exports by consulting FRED.

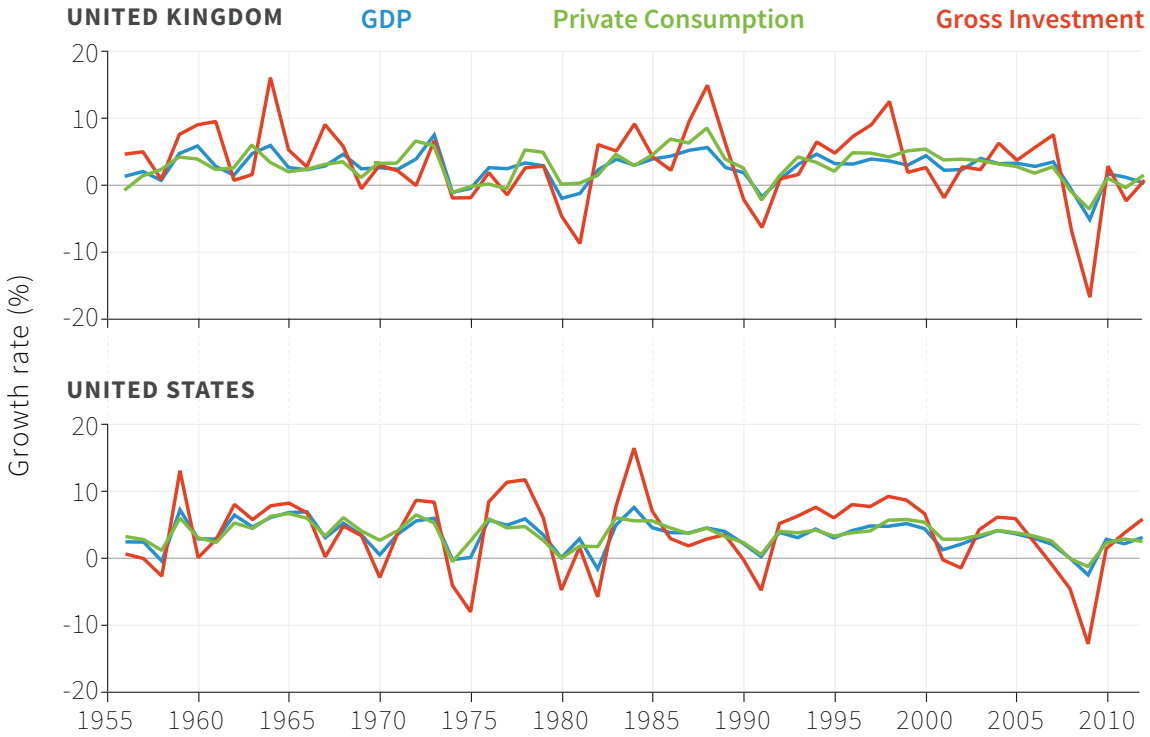


Figure 12.18a *Growth rates of consumption, investment and GDP in the UK and US, percent per annum (1956-2012).*

Source: *Federal Reserve Bank of St. Louis. 2015. 'FRED.'*

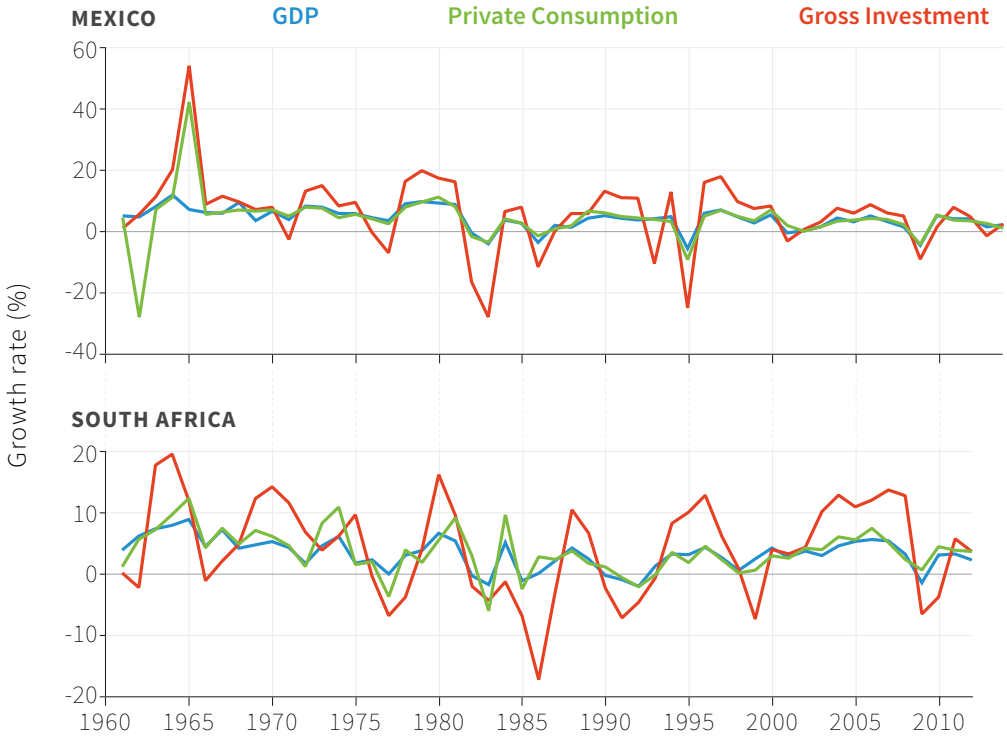


Figure 12.18b *Growth rates of consumption, investment and GDP in Mexico and South Africa (1961-2012).*

Source: *The World Bank. 2015. 'World Development Indicators.' http://data.worldbank.org/data-catalog/world-development-indicators. OECD. 2015. 'OECD Statistics.' http://stats.oecd.org/.*

DISCUSS 12.8: CONSULTING FRED

Use data from FRED to construct, for your own country, charts for the growth rate of real GDP, consumption, investment, net exports and public expenditure.

1. How has public expenditure evolved in your own country since 1960?
2. Comment on the relationship between the growth rate of output and public spending during this period.
3. Describe the volatility of government spending and net exports relative to that of GDP and explain what lies behind the patterns you observe.

12.9 MEASURING THE ECONOMY: INFLATION

In Figures 12.19a and 12.19b we repeat the graphs from Figure 12.3, showing the growth rate of GDP and the unemployment rate in the UK from 1875 to 2014.

Beneath this, in Figure 12.19c, we show the rate of inflation over this period. *Inflation* is an increase in the general price level in the economy, usually measured over a year. For the British economy, inflation ranges from a low level, with prices actually falling—called *deflation*—for much of the inter-war period before and after the Great Depression, to a peak of nearly 25% per annum in 1975.

Previously we saw that the downward spikes of economic crises were associated with upward spikes of unemployment; we now see that inflation was especially low in the 1930s and especially high in the 1970s. The peak in inflation followed the first of two oil price shocks (1973 and 1979) that were major disturbances to the global economy in the 1970s.

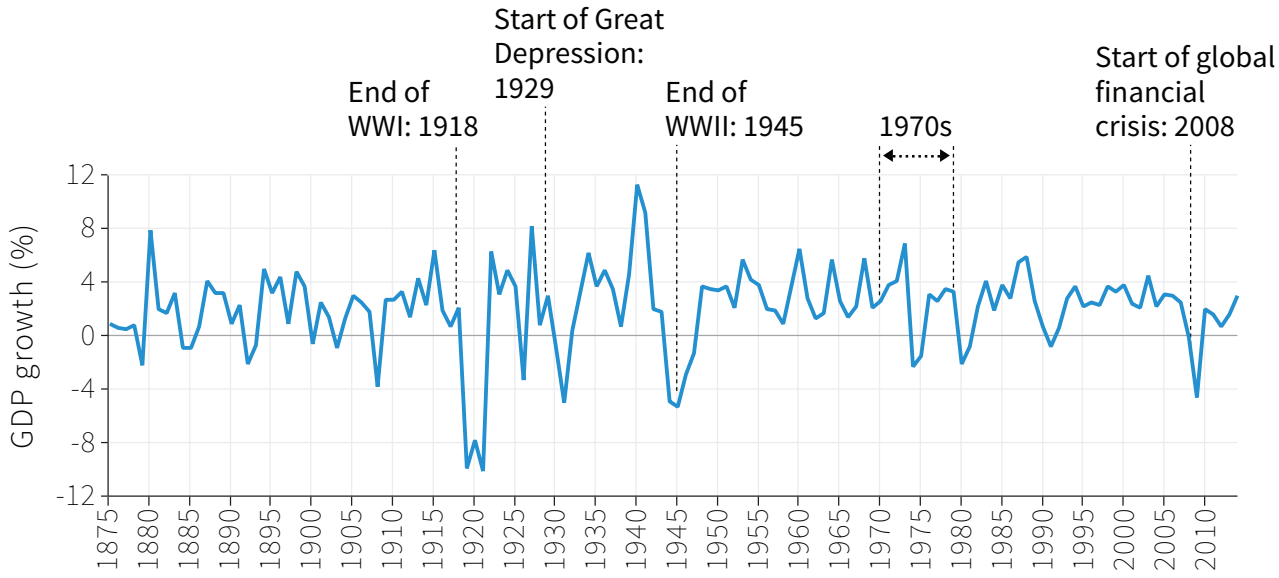


Figure 12.19a *UK GDP growth (1875-2014).*

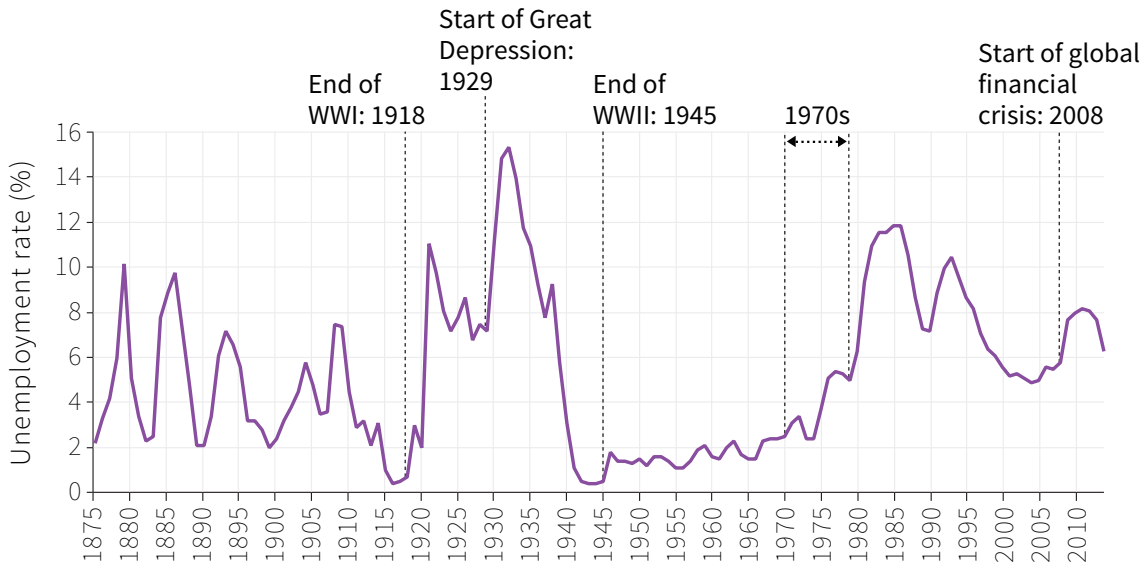


Figure 12.19b *UK unemployment rate (1875-2014).*

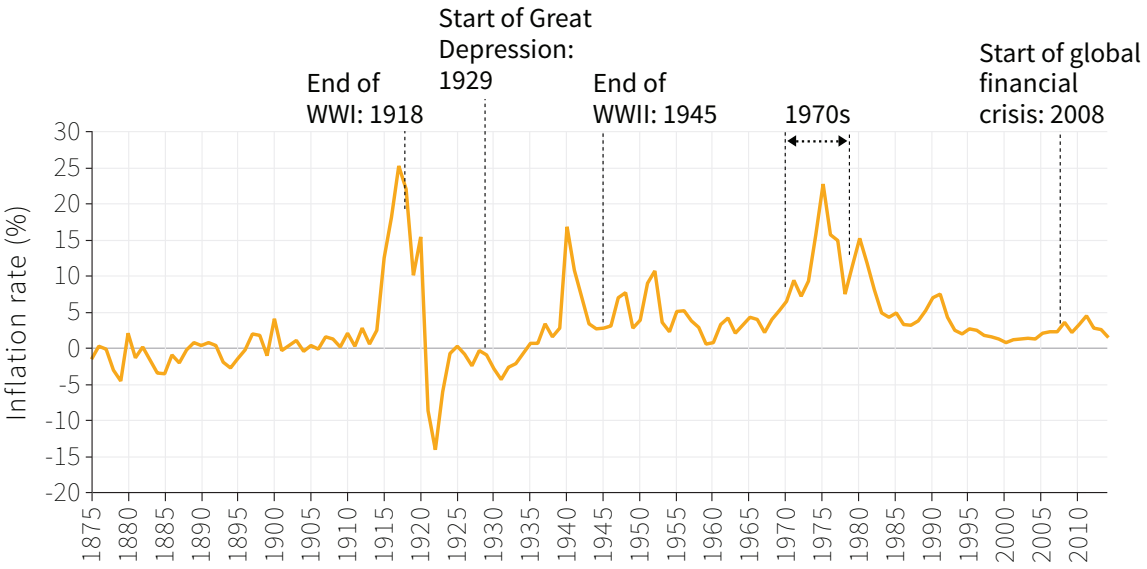


Figure 12.19c *UK inflation rate (1875-2014).*

Source: *Bank of England. 2015. 'Three Centuries of Macroeconomic Data.'*

Figure 12.20 shows average rates of inflation in different regions of the world, and how they have changed over time. Upward spikes in inflation have tended to occur in periods of economic crisis, but the general trend worldwide since the 1970s has been a decline in inflation rates. The figure also shows that inflation tends to be higher in poor than in rich countries. For instance, since 2000 inflation has averaged 6.0% in sub-Saharan African and 6.6% in south Asia, in contrast to only 2.2% in the high-income OECD countries.

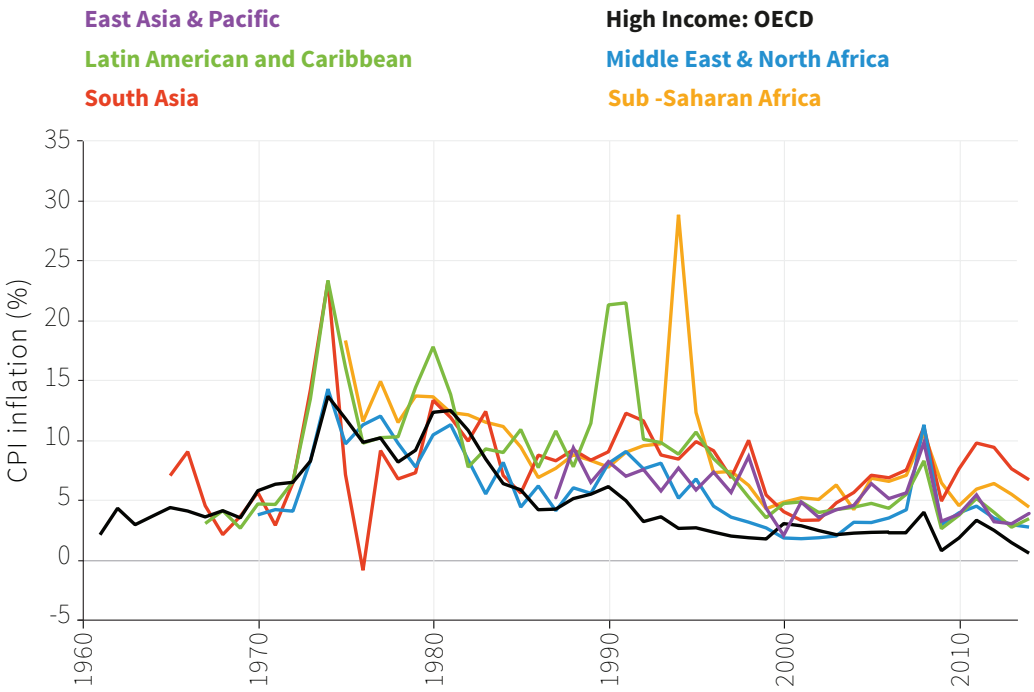


Figure 12.20 *Inflation levels and volatility in high- and low-income economies.*

Source: *The World Bank. 2015. 'World Development Indicators.'*

What is inflation?

Take your favourite chocolate bar. If its price goes up during the year from 50p to 55p, how do you know that is a symptom of inflation in the economy? It could just be that the chocolate bar has become more expensive relative to everything else as a result of a rightward shift in the demand curve or a leftward shift in the supply curve of the kind we studied in Unit 8. To see what has happened to prices across the economy, take a giant shopping basket and fill it with every product and service that you buy in January. Has the price of this same giant basket increased when you check the prices in January the following year? And what about the baskets of other people?

To answer this question, and to understand how inflation is measured, it's best to listen to the people who work it out. In the UK the Office for National Statistics does this. Richard Campbell is the head of the team in charge of measuring inflation, and he has made this animation to explain how the task is done.

DISCUSS 12.9: MEASURING INFLATION

After watching Richard Campbell's animation, answer these questions:

1. How do we construct a giant representative shopping basket for the whole population?
2. If you hear that inflation is 2.5%, this means that the price of the representative shopping basket has increased from £100 to how much?

The official national inflation rate does not necessarily reflect your own personal inflation rate. If you want to calculate your own personal inflation rate and how it deviates from the national one, the UK Office of National Statistics has a personal inflation calculator. Your own office of national statistics may also have a personal inflation calculator.

3. Using a personal inflation calculator, calculate your personal inflation rate and comment on how and why it differs from the official inflation rate for your country.

The *Consumer Price Index* (CPI) measures the general level of prices that consumers have to pay for goods and services, including consumption taxes. The basket of goods and services is chosen to reflect the spending of a typical household in the economy. For this reason, the change in the CPI, or CPI inflation, is often considered to measure changes in the "cost of living".

The CPI is based on what consumers actually buy. It includes the prices of food and drink, housing, clothing, transportation, recreation, education, communications, medical care, and other goods and services. The goods and services in the basket are weighted according to the fraction of household spending they account for. The CPI excludes exports, which are consumed by foreign residents, but includes imports, which are consumed by households in the home economy. The change in the CPI over the past year is commonly used as a measure of inflation.

The *GDP deflator* is a price index like the CPI, but it tracks the change in prices of all domestically produced final goods and services. Instead of a basket of goods and services, the GDP deflator tracks the price changes of the components of domestic GDP, that is, of $C + I + G + X - M$. (The GDP deflator includes exports, which are produced by the home economy, but excludes imports, which are produced abroad.)

The GDP deflator can also be expressed as the ratio of nominal (or current price) GDP to real (or constant price) GDP. The GDP deflator series is most commonly used to transform a nominal GDP series into a real GDP series. As we saw in Unit 1.2 and

Unit 1's Einstein section, the real GDP series shows how the size of the home economy changes over time, taking into account changes in the price of domestically produced goods and services.

DISCUSS 12.10: THE CPI AND THE GDP DEFLATOR

1. Use the data from FRED to construct charts for real GDP growth, the unemployment rate and the inflation rate for the US. Select the period from 1960 until 2014. In addition, download the data for the US GDP deflator (search for GDPDEF).

Use the data you downloaded to answer the following questions (remember that the CPI is calculated from the price of goods consumed in the home country, while the GDP deflator is calculated from the price of the goods produced in the home country):

2. The main difference in the evolution of the series for the CPI and the GDP deflator takes place in 1974-75 and 1979-2000. Why? (Hint: think about the impact of an oil crisis on the price of imported goods and in particular on your own transport and fuel bills.)
3. How did the inflation rates based on the CPI and on the GDP deflator evolve during the 1970s? And since the early 1980s?
4. What do you notice about the evolution of unemployment and inflation in the early 1980s?
5. Now construct the same charts for your own country. Write a brief report on the evolution of inflation, unemployment and real GDP growth rate during the period.

12.10 CONCLUSION

In this unit, we have introduced two essential tools for understanding the economy: the national accounts used to measure aggregate economic activity, and a set of models that allow us to organise the data in ways that illuminate economic fluctuations. Economists are often asked to provide forecasts about the future development of the economy and they use both data and models to do this. We have learnt in this unit that households and firms make forecasts when deciding on their spending.

In the following two units, we focus on government policy. We shall see that the government and central bank need to take into account how households and firms think about the future and what may frustrate their plans, if they are to make good forecasts and good policy.

CONCEPTS INTRODUCED IN UNIT 12

Before you move on, review these definitions:

- *Recession*
- *Participation rate, Unemployment rate, Employment rate*
- *Okun's law*
- *Circular flow of income and spending*
- *Aggregate demand* and its components: *C, I, G, X, M*
- Government *transfer payments*
- *Consumption smoothing*
- *Self-insurance* and *Co-insurance*
- *Capacity utilisation rate*
- Investment as a *coordination game*
- *Inflation, CPI* and *GDP deflator*

Key points in Unit 12

Shocks

The economy is continuously hit by shocks such as investment booms based on new technology, unusually good or bad weather, war, disease, loss of confidence, and oil price shocks. These shocks result in large unexpected changes in income, employment and prices.

National accounts

We can use the national accounts to measure the size of the economy and fluctuations in total output (GDP) and its constituent components: consumption, investment, government spending, exports and imports.

Smoothing consumption

Throughout history and across different types of economy, through forward-looking behaviour, households seek to minimise the fluctuations in their access to goods and services. They do this by borrowing, saving and sharing. Governments smooth consumption through unemployment benefits.

Limits to smoothing

When incomes fluctuate, families do not entirely eliminate fluctuations in their consumption. The reason is that self-insurance is limited by credit constraints and weakness of will; while co-insurance is limited by the lack of government programs, and the modest access that most families have to transfers from family and friends.

Volatile investment

Investment is more volatile than consumption because an investment is easily postponed while consumption is not, the clustering of investment around new technologies and the role of beliefs in the vicious and virtuous circles of investment.

12.11 EINSTEIN

Ratio scales and logarithms

In Unit 1, we made frequent use of a ratio or log scale on the vertical axis to display long-run data. For example, we used ratio scales with the units doubling in Figure 1.1b and rising tenfold in Figure 1.2. The ratio scale is also called a logarithmic, or log, scale. We can write a scale where the tick marks on the vertical axis double like this:

$$2^0, 2^1, 2^2, \ldots$$

Or a scale where they rise tenfold, like this:

$$10^0, 10^1, 10^2, \ldots$$

The first is called a logarithmic scale in base 2; the second is in base 10.

As we saw in the charts in Unit 1, if the data forms a straight line on a ratio (logarithmic) scale, then the rate of growth is constant. A different method of using this property of logarithms is to first convert the data into natural logs and then plot it on a scale that is linear in logs. Natural logs use base e.

We can use a calculator or a spreadsheet programme to convert levels into natural logs: as you can see, when applied to this data, it converts the curved line in Figure 12.2 in the left-hand panel into one that is almost a straight line in the right-hand one.

Using the chart functions in Microsoft Excel helps illustrate the relationship between plotting the data with a ratio scale on the vertical axis (Figure 12.21a, which uses the doubling or base 2 scale) and transforming the data into natural logs and plotting on a linear scale (in logs) on the axis (Figure 12.21b). Note that the tick marks double from 4,096 to 8,192 to 16,384 in Figure 12.21a and rise from 8.5 to 9 to 9.5 in Figure 12.21b.

In each chart, a line appears alongside the data series. Using Excel, we created Figure 12.21a by selecting Analysis/Trendline, and selecting “Exponential”. Excel finds the line or curve that best fits the data points: since the scale is a ratio scale, a straight line is displayed. The equation of the line is given. Other spreadsheet or graphing software offers similar features.

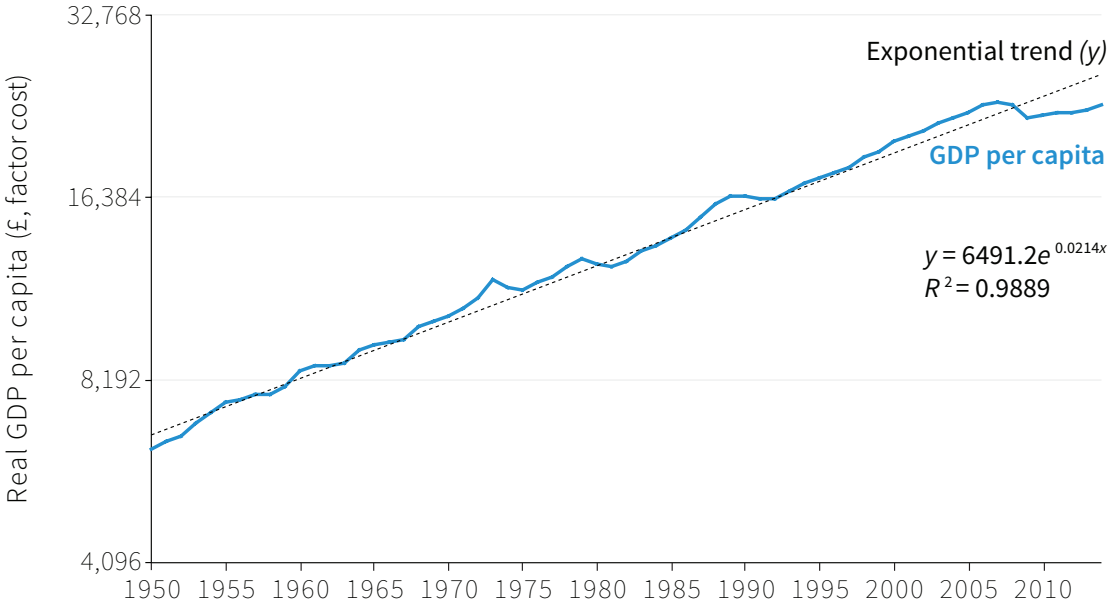


Figure 12.21a *Ratio scale and exponential function.*

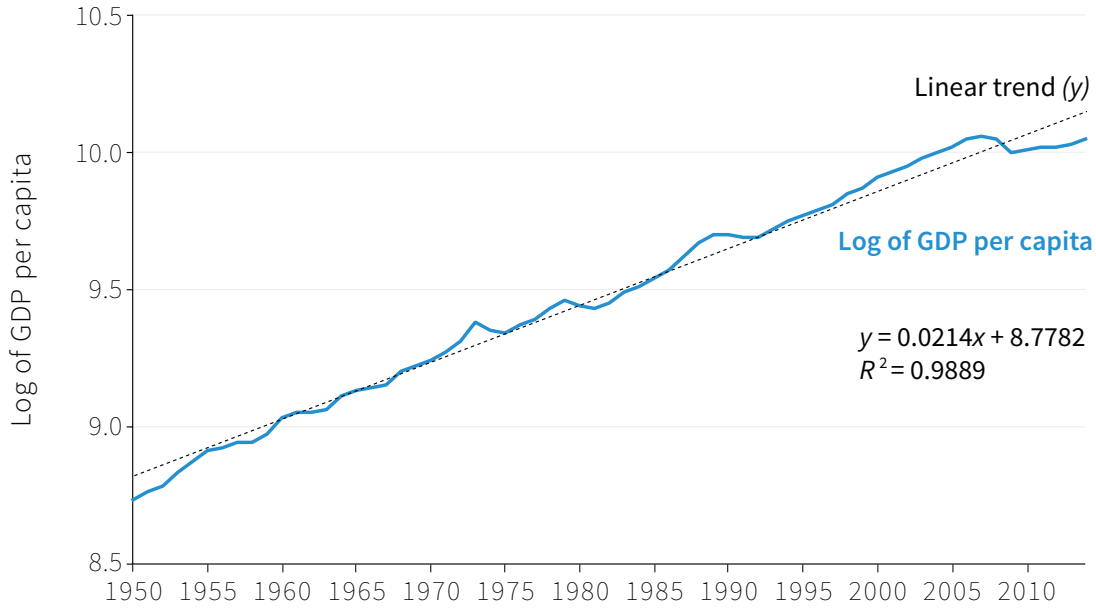


Figure 12.21b *Linear scale in natural logs and linear function.*

We can see that the exponential function uses what is called base *e* in contrast to base 2 (doubling) or base 10 (increasing tenfold). The exponent on *e* tells us the exponential, or equivalently, the compound annual growth rate of the series: it is *0.0214 x 100* = *2.14%* per annum.

In Figure 12.21b, if we use Excel to select the "Fit a linear function" option, a straight line appears: this time, we see an equation for a straight line with intercept 8.7782 and slope 0.0214. The slope of the line tells us the exponential, or equivalently, the compound annual growth rate of the series: it is *0.0214 x 100* = *2.14%* per annum.

In summary:

- When a data series is plotted, either using a ratio scale or by transforming the data into natural logs, *and the outcome is approximately linear*, the growth rate of the series is constant. This constant growth rate is called an exponential growth rate.
- The exponential growth rate (known also as the compound annual growth rate or CAGR) is *the slope of the line when the natural logarithm of the data series is plotted.*
- Notice the persistent deviation of the British economy from the trend line following the 2008 financial crisis.

Okun's law

This is defined as:

$$\Delta u_t = \alpha + \beta(GDP\ growth)_t$$

where Δu_t is the change in unemployment rate at time t, $(GDP\ growth)_t$ is the real GDP growth at time t, α is the intercept value, and β is a coefficient determining how real GDP growth is predicted to be translated into a change in unemployment rate. Okun's is an empirical linear relationship that associates real GDP growth with changes in unemployment. The coefficient β is generally found to be negative, suggesting that a positive real GDP growth will be associated with a fall in the unemployment rate.

The estimated Okun's law relationship for Germany, for the period 1970-2011, has coefficients β = - 0.21 and α = 0.57.

When we estimate a line of best fit, we also measure R^2, which is a statistic that lies between 0 and 1. It measures how closely the numbers we observe fit the line that we draw through them, with 1 being a perfect fit, and 0 representing no observable relationship between the observations and the prediction. In our case, the statistic measures how well Okun's law approximates the data for real GDP growth and unemployment changes. The R^2 statistic is 0.24 for Germany for the period 1970-2011, which is much lower than for the estimated Okun's law equation for the US, which is 0.69.

To work out the predicted percentage change in unemployment for Germany in 2009 using the Okun's law equation, we simply plug in the value of real GDP growth for Germany in 2009 and solve the equation as follows:

$$\Delta u_{2009} = 0.57 + 0.21 \times (-5.1) = 1.65$$

Okun's law predicts that the fall in GDP of 5.1% in 2009 in Germany should have been associated with an increase in unemployment by 1.65 percentage points.

12.12 READ MORE

Bibliography

1. Algan, Yann, Elizabeth Beasley, Florian Guyot, and Fabrice Murtin. 2014. 'Big Data Measures of Human Well-Being: Evidence from a Google Stress Index on US States.' *Sciences Po Working Paper.*
2. Bank of England. 2015. 'Three Centuries of Macroeconomic Data.'
3. Broadberry, Stephen, Bruce M. S. Campbell, and Alexander Klein. 2015. *British Economic Growth, 1270-1870*. Cambridge: Cambridge University Press.
4. Buiter, Willem, and Ebrahim Rahbari. 2015. *What Is A (Global) Recession?* Citi Research Global Economics View.
5. Clark, Andrew E, and Andrew J Oswald. 2002. 'A Simple Statistical Method for Measuring How Life Events Affect Happiness.' *International Journal of Epidemiology* 31 (6): 1139–44.
6. Collins, Daryl, Jonathan Morduch, Stuart Rutherford, and Orlanda Ruthven. 2009. 'Portfolios of the Poor.'
7. Coyle, Diane. 2014. *GDP: A Brief but Affectionate History*. Princeton, NJ: Princeton University Press.
8. Durante, Ruben. 2010. 'Risk, Cooperation and the Economic Origins of Social Trust: An Empirical Investigation.' *Sciences Po Working Paper*.
9. Eurostat. 2015. 'Confidence Indicators by Sector.'
10. Federal Reserve Bank of St. Louis. 2015. 'FRED.'
11. Fletcher, James. 2014. 'Spurious Correlations: Margarine Linked to Divorce?' *BBC Magazine*, May 26.
12. International Labour Association. 2015. 'ILOSTAT Database.'
13. Jappelli, Tullio, and Luigi Pistaferri. 2010. 'The Consumption Response to Income Changes.' *VoxEU.org*.
14. Naef, Michael, and Jürgen Schupp. 2009. 'Measuring Trust: Experiments and Surveys in Contrast and Combination.' *IZA Discussion Paper No. 4087*.
15. OECD. 2010. 'Employment Outlook 2010: Moving beyond the Jobs Crisis.'
16. OECD. 2015. 'OECD Statistics.'
17. Read, Daniel, and Barbara van Leeuwen. 1998. 'Predicting Hunger: The Effects of Appetite and Delay on Choice.' *Organizational Behavior and Human Decision Processes* 76 (2): 189–205.
18. *The Economist*. 2009. 'Smooth Operators.' May 14.
19. *The Economist*. 2012. 'New Cradles to Graves.' September 8.
20. The World Bank. 2015. 'World Development Indicators.'
21. US Bureau of Economic Analysis. 2015. 'Fixed Assets Accounts Tables.'

Beta December 2015 version

13

UNEMPLOYMENT AND FISCAL POLICY

HOW GOVERNMENTS CAN MODERATE COSTLY FLUCTUATIONS IN EMPLOYMENT AND INCOME

- Fluctuations in aggregate demand affect GDP through a multiplier process, because households face limits to their ability to save, borrow and share risks
- An increase in the size of government following the second world war coincided with smaller economic fluctuations
- Governments can use fiscal policy to stabilise the economy, but bad policy can destabilise it
- If a single household saves, its wealth necessarily increases; if all households save this may not be true because, without additional spending by the government or firms to counteract the fall in demand, income will fall
- Every economy is embedded in the world economy. This is a source of shocks, both good and bad, and places constraints on the kinds of policies that can be effective

Institute for New Economic Thinking
UCL SciencesPo Azim Premji University
FRIENDS PROVIDENT Foundation

In August 1960, three months before he was elected US president, the 43-year-old Senator John F. Kennedy found time to spend the day cruising Nantucket Sound on his boat, the *Marlin*. His crew for the day included John Kenneth Galbraith and Seymour Harris, both Harvard economists, and Paul Samuelson, an economist at MIT and later also a Nobel laureate. They had not been recruited for their nautical skills. In fact, the senator did not even know them.

The future president wanted to learn "the new economics" which John Maynard Keynes, an economist who we will learn more about in section 13.6, had formulated in response to the Great Depression. When Kennedy was a teenager in the decade before the second world war, the US and many other countries of the world experienced a drastic fall in output (we can see this for the US in Figure 13.1) and massive unemployment that persisted for more than 10 years.

Kennedy had a lot to learn: he admitted that he had barely passed the only economics course he took at Harvard. He would later spend a day at the America's Cup sailing races being tutored by Harris, who assigned texts for him to read. Harris later gave private lessons to the senator, shuttling by air between Boston, where he worked, and Washington DC.

In 1948, Samuelson had written *Economics*, the first major textbook to teach these new ideas. Harris promoted the same economic ideas in a book that he edited in 1948, a collection of 31 essays by 24 contributors called *Saving American Capitalism*. It seemed at that time that capitalism needed saving: a model promoted as the alternative to capitalism, the centrally planned economies of the Soviet Union and its allies, had entirely avoided the Great Depression. Kennedy needed economics to understand policies to promote economic growth, reduce unemployment, but also avoid economic instability.

We have seen in Unit 12 that instability in the economy as a whole is characteristic not only of economies dominated by agriculture, but also of capitalist economies. Figure 13.1 shows the annual growth of real GDP in the US economy since 1870.

A dramatic reduction in the severity of business cycles occurred after the end of the second world war. Figure 13.1 shows another important development at that time: the increasing role of government in the economy. The red line shows the share of federal (national), local and state government tax revenue as a share of GDP. This is a good measure of the size of the government sector in the economy.

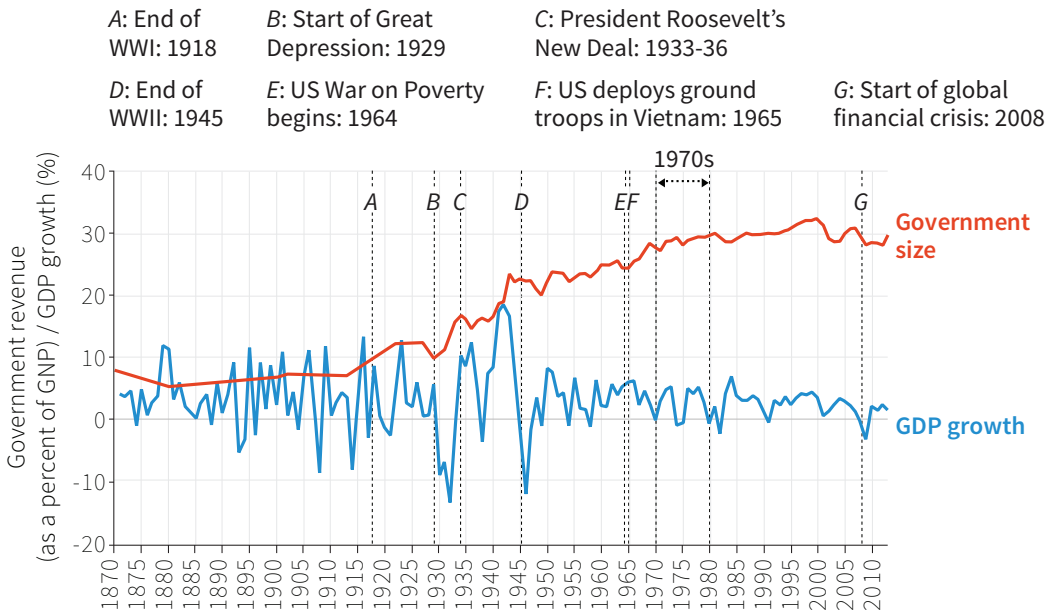


Figure 13.1 *Fluctuations of output and the size of government in the US (1870-2013).*

Source: *The Maddison Project. 2013. '2013 Version.'; US Bureau of Economic Analysis. 2015. 'GDP & Personal Income.'; Wallis, John Joseph. 2000. 'American Government Finance in the Long Run: 1790 to 1990.' Journal of Economic Perspectives 14 (1): 61–82.*

The share of employment in agriculture, which we have seen was one cause of volatility in the economy, fell from 44% in the 1870s to 18% by the beginning of the second world war, yet there was no sign of the economy becoming more stable over this period. As we have seen, households try to smooth fluctuations in their consumption but, in part because there are limits to how much they can borrow, they can't always do this successfully.

The fact that the fluctuations in output growth are dramatically reduced as the size of government expands does not demonstrate that increased government spending stabilised the economy. (Remember: statistical correlations do not demonstrate causation.) But there are good reasons to think that the increase in the red line may have been part of the cause of the smoothing of the blue line. In this unit we ask why the increased role of the government in the economy is part of the explanation of the more stable economy in the second half of the 20th century.

What Harris taught Kennedy was influenced by the contrast between the volatility of the economy before the second world war, and the steadier growth and absence of deep recessions afterwards. Why do economies experience unemployment, inflation and instability in output, and what kinds of policies might address these problems?

In Unit 12, we took the household's eye-view of the business cycle, which allowed us to establish why fluctuations in employment and income are costly, and how households try to limit the consequences for wellbeing. In this unit, we take the policymaker's viewpoint. As we saw in Figure 13.1, the big increase in the size of government after the second world war was accompanied by a reduction in the size of

business cycle fluctuations. After 1990, the business cycle in the advanced economies became even smoother, until the global financial crisis in 2008. This led to the period from the early 1990s to the late 2000s being called the *great moderation*.

13.1 THE TRANSMISSION OF SHOCKS: THE MULTIPLIER MECHANISM

In a capitalist economy, private investment spending is driven by expectations about future post-tax profits. As we saw in Unit 12, spending on investment projects tends to cluster. Two reasons for this:

- Firms may adopt a new technology at the same time.
- Firms may have similar beliefs about expected future demand.

We need a tool to help us understand how decisions of firms (and households) to raise or reduce investment spending will affect the economy as a whole. You will recall that:

- Households that are able to completely smooth the bumps in their income do not respond with higher consumption to the higher employment that comes with a temporary investment boom.
- But, in credit-constrained households, higher income from getting a job or moving from part-time to full-time work will also lead to higher consumption spending.

As a result, changes in current income influence spending, affecting the income of others, so indirect effects through the economy amplify the direct effect of a shock to *aggregate demand* (often shortened to AD) created by an investment boom.

We will show how economists answer such questions as "how large would the total direct and indirect impact of a rise in investment spending be?" or, conversely, "what would be the effect of lower government spending?"

A statistic called *the multiplier* provides one way of answering this question. Imagine there is a new technology. New spending takes place in the economy as a result; output of the new capital goods rises, as do the incomes of the people producing them. The circular flow of expenditure, income and output shown in Figure 12.7 illustrates this process.

- *If the increase in GDP is equal to the initial increase in spending:* We say that the multiplier is equal to one.

- *If the total increase in GDP is greater than the initial increase in spending:* We say that the multiplier is greater than one.

To see why GDP may rise by more than the initial increase in investment spending, we explain what economists call the *multiplier* process. We do this by combining the very different behaviour of consumption-smoothing and non-smoothing households to represent consumption spending for the economy as a whole. In this *aggregate consumption function*, consumption depends on current income among other things. Recall that in the model of Unit 12 consumption-smoothing households will not increase their consumption one-for-one, or even at all, in response to a temporary €1 increase in their income. Credit-constrained and other households who do not smooth, on the other hand, will increase their consumption by €1 in response to a temporary €1 increase in their income.

In 2008, when governments considered temporary increases in government spending and cuts in taxes in response to the recession that followed the global financial crisis, the size of the multiplier became the subject of a debate among policymakers and in economics blogs. We return to the debate later in the unit.

As we shall see, in an aggregate consumption function in which spending resulting from a temporary €1 increase in income is greater than zero but less than €1 (say, for example, 60 cents), the multiplier is greater than one.

After explaining how this is a consequence of the multiplier process, we will consider the assumptions that we make in the model. We also show that the validity of the assumptions we make in the multiplier model depends on the state of the economy.

13.2 THE MULTIPLIER MODEL

We begin with a simple model that excludes the government and foreign trade. In this model, there are two types of expenditure:

- *Consumption*
- *Investment*

Aggregate consumption spending is assumed to have two parts:

- *A fixed amount:* How much one will spend even with no income. The fixed amount is shown as c_0 on the vertical axis of Figure 13.2.
- *A variable amount:* This depends on current income, and is an upward-sloping red line in Figure 13.2.

So we can write consumption spending in the form of an equation. This is the *aggregate consumption function* that we introduced in the previous section:

$$\begin{aligned} \textit{aggregate consumption} &= \textit{autonomous consumption} \\ &\quad + \textit{consumption that depends on income} \\ C &= c_0 + c_1 Y \end{aligned}$$

The term c_1 gives the effect of one additional unit of income on consumption, called the *marginal propensity to consume*. In Figure 13.2, the slope of the consumption line is equal to the marginal propensity to consume. A steeper consumption line means a larger consumption response to a change in income. A flatter line means that households are smoothing their consumption so that it does not vary much when their incomes change. We assume that the marginal propensity to consume is less than one.

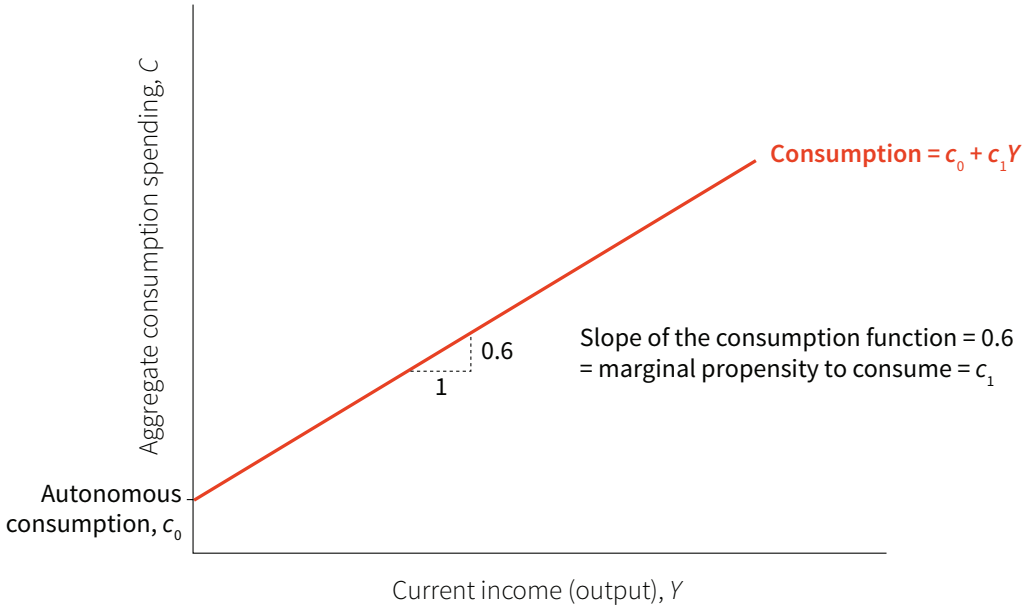


Figure 13.2 *The aggregate consumption function.*

We shall work with an aggregate consumption function in which the marginal propensity to consume, c_1, equals 0.6. This means that an additional euro of income increases consumption by *€1 x 0.6 = 60 cents*. The marginal propensity to consume is positive, but less than one: this says that only part of an increase in income is consumed; the rest is saved.

Naturally, this average number hides large variation across households, which differ in their wealth and in the credit constraints they face. Most households have little wealth, and even in rich countries about one in four are credit-constrained. As we saw in Unit 12, weakness of will also plays a role. So, both for households that are credit-constrained and for those that do not save ahead of anticipated declines in income, consumption tracks income closely.

Households with low wealth smooth consumption very little if their income falls sharply. The marginal propensity to consume for this group is closer to 0.8. For the small fraction of households who hold the majority of wealth, however, current income plays a very small role in determining consumption, and their marginal propensity to consume is closer to zero. This means that for rich households, an increase in current income of €1 would raise their consumption by just a few cents.

The term c_0 in the aggregate consumption function is a "catch-all" for all the other influences on consumption that are not related to current income. Taken literally, as you have seen, it is how much a person with no income would consume; but this is not the best way to think about it. It is just the consumption that is independent of income, and for this reason we call it *autonomous consumption*.

Since only current income is included explicitly in the consumption function, expectations about future income will be included in autonomous consumption. To see what this means in practice, recall from Unit 12 that consumption will change as a result of people becoming more or less optimistic about their future employment and earnings prospects.

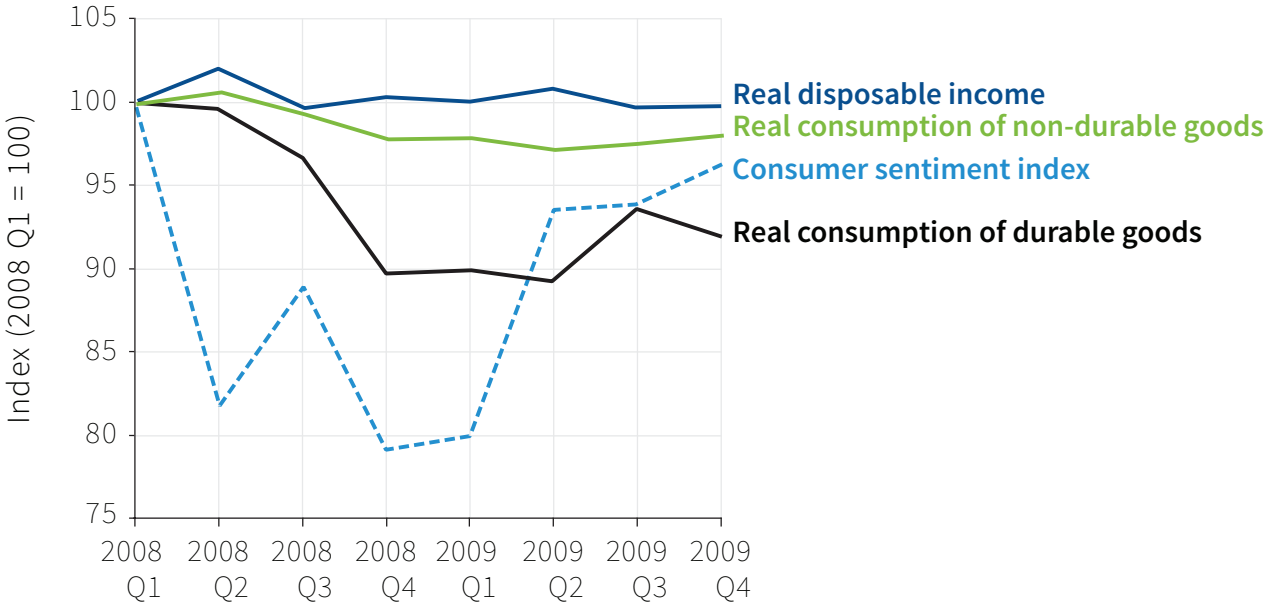


Figure 13.3 *Fear and household consumption in the US during the global financial crisis (Q1 2008 to Q4 2009).*

Source: *Federal Reserve Bank of St. Louis. 2015. 'FRED.'*

Figure 13.3 illustrates how expectations affected consumption in the financial crisis of 2008 and highlights the exceptional nature of this episode. The figure shows how consumer confidence changed in the US over the course of the crisis. The consumer sentiment index that we have used is the University of Michigan Surveys of Consumers. It is based on monthly interviews with 500 households, and asks how they view prospects for their own financial situation and for the general economy over the short and long term. The figure also plots the evolution of a number of key

macroeconomic indicators: disposable income, consumption of durable goods like cars and home furnishings, and consumption of nondurable goods, such as food. All of the series in Figure 13.3 are shown as index numbers, with the first quarter of 2008 as the base year.

We notice:

- *Consumption of nondurable goods went down slightly more than disposable income:* It fell by 3% during the period. Contrary to the predictions of consumption smoothing, households were sufficiently worried about their future prospects that they made adjustments to their spending on nondurables.
- *Consumption of durables decreased much more dramatically than disposable income:* By 10%.

Why the sudden drop in consumption of *consumer durables*? An important reason is that households were suddenly fearful about the future of their jobs, as shown by the sharp downturn in the consumer sentiment index in Figure 13.3. The collapse of the investment bank Lehman Brothers in September 2008, worries about the stability of the banking system, and a higher burden of household debt due to falling house prices led households with mortgages to postpone purchases of expensive items like cars and fridges. It's important to remember that spending on consumer durables can easily be postponed: in this sense it is more like an investment than a consumption decision (even though consumer durables are counted as part of consumption in the national accounts). As a result, we would expect the series for consumer durables to be more volatile than for nondurable consumption.

We now show how a shock is transmitted through the economy. In Figure 13.4 we show the amount of output produced by the economy (on the horizontal axis) and the demand for output (on the vertical axis). Everything is measured in real terms because we are interested in how changes in aggregate demand create changes in output and employment.

The 45-degree line from the origin of the diagram shows all the combinations in which output is equal to aggregate demand. This corresponds to the circular flow discussed in Unit 12, where we saw that spending on goods and services in the economy (aggregate demand) is equal to production of goods and services in the economy (output). You can see this because with a 45-degree line the horizontal distance (output) is equal to the vertical distance (aggregate demand). Point A in Figure 13.4 shows an output–aggregate demand combination where output and aggregate demand are equal. Point A is called a *goods market equilibrium*: the economy will continue producing at that output level unless something changes spending behaviour. We can therefore say that:

$$\textit{output} = \textit{aggregate demand for goods produced in the home economy}$$
$$Y = AD$$

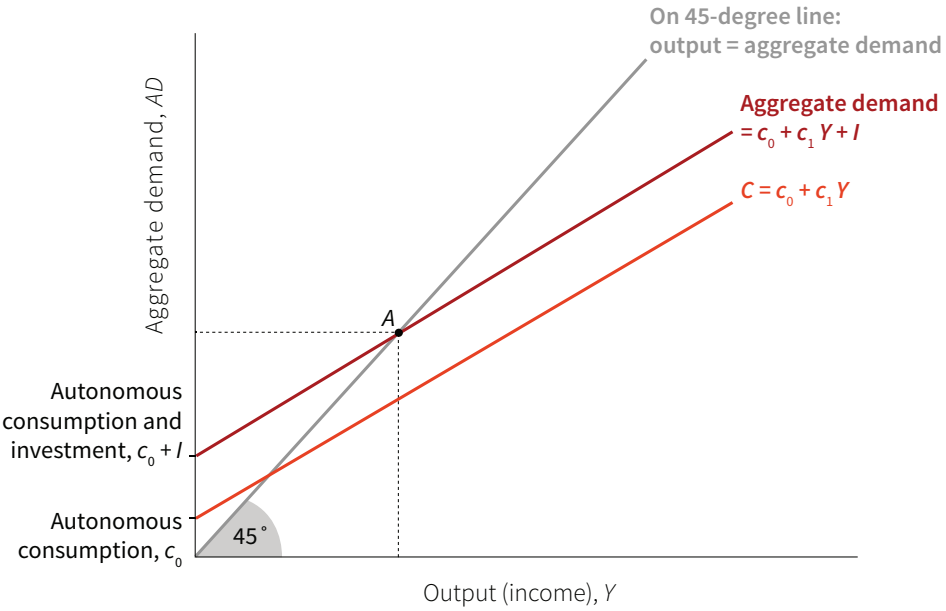


Figure 13.4 *Goods market equilibrium: The multiplier diagram.*

But how do we know where the economy is on the 45-degree line? Is it at a position of low output, which would mean high unemployment, or is it at a position of high output, which would mean low unemployment?

In this model, we assume that the level of output depends on the level of aggregate demand, which we determine by analysing its individual components. We assume that firms are willing to supply any amount of the goods demanded by those making purchases in the economy.

To find out the output of the economy located in the country (economists call this *home*'s output level) we need to add together the elements of the spending of the groups that purchase the goods and services produced by the home economy. Because we have assumed that there is no government spending and trade with other economies, there are just two components of aggregate spending:

- *Consumption*: To include this we take the consumption line introduced in Figure 13.2. Because the propensity to consume is less than one, the consumption line is flatter than the 45-degree line, which has a slope of one.
- *Investment*: We assume it does not depend on the level of output.

The equation for aggregate demand is therefore:

$$aggregate\ demand = consumption + investment$$
$$AD = C + I$$
$$AD = c_0 + c_1 Y + I$$

So adding investment to the consumption line simply leads to a parallel upward shift. In this respect investment is similar to autonomous consumption. We can see from Figure 13.4 that the aggregate demand line has an intercept of $c_0 + I$, a slope of c_1, and is flatter than the 45-degree line.

In Figure 13.4, we now have a picture showing how the level of output in the economy is determined. The same figure tells us the effect of a change in autonomous consumption, c_0, or in investment. We study this change exactly as we did the changes in supply and demand in Unit 9: we see how the change makes the old outcome no longer an equilibrium, and then we locate the new equilibrium. The expected change is the movement from the old to the new equilibrium.

Changes in autonomous consumption or investment displace the old equilibrium because they change aggregate demand, which in turn alters the level of output and employment. In Figure 13.5, we take the multiplier diagram and reduce investment. We choose a reduction in investment of €1.5bn. Click through Figure 13.5 to see what happens.

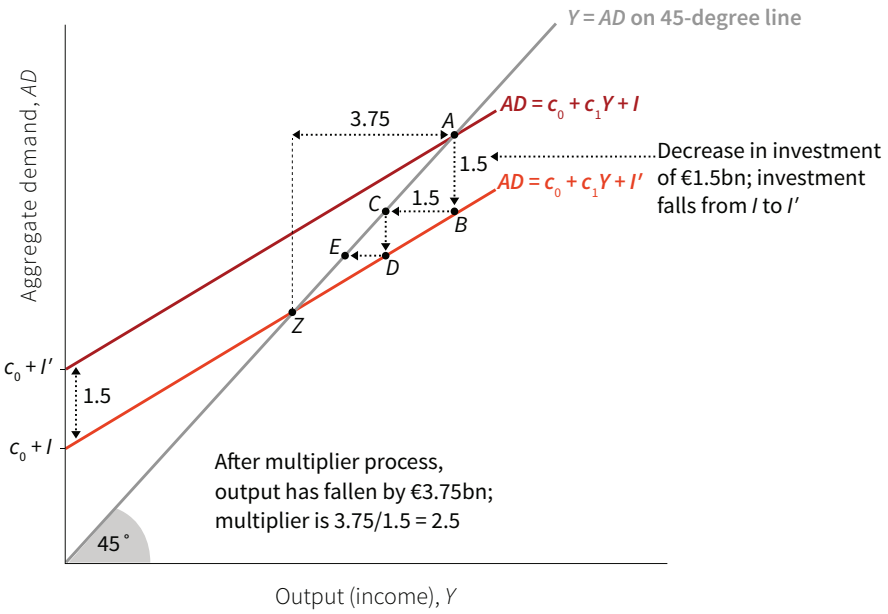


Figure 13.5 *The multiplier in action: Investment-led recession.*

We trace the effect of the fall in investment through the economy in Figure 13.5. The fall in investment cuts aggregate demand by €1.5bn, and the economy moves vertically downward from point *A* to point *B*. Firms cut back production, and with output and employment lower, incomes fall by €1.5bn. But this is just the first round.

In the second round, if you look again at the consumption equation, you will see that a fall in income in turn leads to lower spending by households. Think of credit-constrained households where someone loses their job: they would like to keep consumption stable but when their income falls, they are unable to borrow enough to sustain that level of consumption, and they reduce their spending. The consumption

equation tells us that this kind of behaviour leads to a fall in aggregate consumption of 0.6 times the fall in income: this is the distance from point *C* to point *D*. The process will go on until the economy reaches point *Z*.

Following the investment shock, the intercept of the line has moved down by €1.5bn, causing a parallel shift in the aggregate demand line. Output has fallen by €3.75bn, more than the fall in investment of €1.5bn: this is the *multiplier process*.

In this case, the multiplier is equal to 2.5, because the total change in output is 2.5 times larger than the initial change in investment. A multiplier of 2.5 is unrealistically large. As we shall see in the next section, once taxes and imports are introduced in the model, the multiplier shrinks.

We call the model of aggregate demand that includes the multiplier process the *multiplier model*. This is a summary:

- *A fall in demand leads to a fall in production and an equivalent fall in income:* This leads to a further (smaller) fall in demand, which leads to a further fall in production, and so on.
- *The multiplier is the sum of all these successive decreases in production:* Eventually, output has fallen by a larger amount than the initial shift in demand. Output is a multiple of the initial shift. This is why it is called a multiplier.
- *Production adjusts to demand:* Firms supply the amount of goods demanded at the prevailing price. When demand falls, firms adjust production down. The model assumes that they do not adjust their prices.

Note that the economy we are studying is one in which we assume there are underutilised resources in the form of spare capacity in production facilities and underemployed labour. We also assume that wages are not affected by changes in the level of output. For the multiplier process to work in the same way in reverse in response to a rise in investment, the assumption of spare capacity and fixed wages means that costs will not rise when output goes up, so firms will be happy to supply the extra output demanded without adjusting their price.

If the economy is not characterised by spare capacity and constant wages, the multiplier will be smaller than what we find here.

We asked what the effect on the output of the economy would be if either investment or autonomous consumption were to change. To answer this, we do two things:

1. *We determine what the output will be for some given level of c_0 and I:* There is an equilibrium in the goods market at which output is equal to aggregate demand. We use this to determine output (where the *AD* curve crosses the 45-degree line). Since we now have a model of aggregate demand, we can solve the equation to get an expression with output on the left-hand side and all the other variables on the right-hand side.

$$
\begin{aligned}
\textit{output} &= \textit{aggregate demand} \\
\textit{output} &= \textit{consumption} + \textit{investment} \\
Y &= C + I \\
Y &= c_0 + c_1 Y + I \\
Y(1 - c_1) &= c_0 + I \\
Y &= \underbrace{\frac{1}{(1-c_1)}}_{\textit{multiplier}} \times \underbrace{(c_0 + I)}_{\textit{autonomous demand}}
\end{aligned}
$$

2. *We use this information to determine how output will change if either autonomous consumption or investment changes:* The equation answers the question. We can calculate how much output will increase or decrease by the value of the multiplier times the change in *autonomous demand*.

Discover another way to summarise our findings from the diagram algebraically in the Einstein section.

Since the multiplier is greater than one (in this case, it is equal to 2.5), the change in output in Figure 13.5 is 2.5 times greater than the initial shock to investment, which means that the shock has been amplified. In algebra, we write this as $\Delta Y = k\Delta I$, and say it as "*delta Y* (the change in output) is equal to k, the multiplier, times *delta I* (the change in investment)".

A value of the multiplier of 2.5 is higher than estimates based on empirical data. An important reason for this is that, so far, the model does not include the impact of taxation or imports. Both of these will dampen the impact on GDP of the initial rise in spending, as we see next.

13.3 HOUSEHOLD TARGET WEALTH, COLLATERAL AND CONSUMPTION SPENDING

From Unit 12, we know that in most economies consumption is the largest component of GDP. Therefore an important part of understanding changes in output and employment is understanding why consumption changes.

We saw that a shock to investment shifts the aggregate demand curve, and is transmitted through the economy as households adjust their spending in response to changes in income. We focused on incomplete smoothing, such as credit constraints. This behaviour is reflected in the size of the multiplier and the slope of the aggregate demand curve. But consumption and saving behaviour can also *shift* the aggregate demand curve.

A shift in aggregate demand can be caused by a shift in autonomous consumption, represented by the term c0 in the aggregate consumption function, $C = c_0 + c_1Y$. This will in turn produce a multiplier response of output and employment through the circular flow of expenditure, output and income in the same way as the fall in investment in the previous section was multiplied.

Think about a family with a mortgage on its house. If the price of houses falls, the family will be concerned that its wealth, too, has fallen. A likely reaction to this is for the household to save more. This is called *precautionary saving*. This behaviour suggests that households have a notion of their *target wealth*.

When something happens to affect the stock of the household's wealth relative to this target, it reacts by either increasing or decreasing saving to restore it to target. If this involves precautionary saving it is modelled as a fall in autonomous consumption.

TARGET WEALTH

Target wealth is the level of wealth that the household aims to hold, based on its economic goals (or preferences) and expectations. We assume that households try to maintain this level of wealth in the face of changes in their economic situation, as long as it is possible to do so.

In 1929 a downturn in the US business cycle that was, on the face of it, similar to others in the preceding decade turned into a large-scale economic disaster—the Great Depression.

The fall in output and employment during the Great Depression highlights two ways in which aggregate consumption might fall—credit constraints in the multiplier process, and changes in wealth relative to target.

To understand the economic mechanisms at work in the Great Depression, we use the multiplier diagram shown in Figure 13.6. Point *A* shows the initial situation of the economy in the third quarter of 1929. There was then a fall in investment. This shifts the aggregate demand curve from the pre-crisis to the crisis level. The dotted line from point *B* shows the level of output that would have been observed in the business cycle trough had the usual multiplier process been at work. There would have been a recession, but no Great Depression. In 1929 there was a fall in the demand for consumer goods even by those who kept their jobs.

Consumption was cut back through two mechanisms:

- *The shift from A to B:* As output and employment fell, some households cut spending on housing and consumer durables because they were credit-constrained, and therefore unable to borrow in the deteriorating conditions. Some economists have estimated that the size of the multiplier at the time was about 1.8.

- *The shift from B to C:* Even households that remained in work cut back spending because it became increasingly clear that the downturn was the new reality, not a temporary shock. This shifted the consumption function down and pulled the economy further into depression from B to C in Figure 13.6.

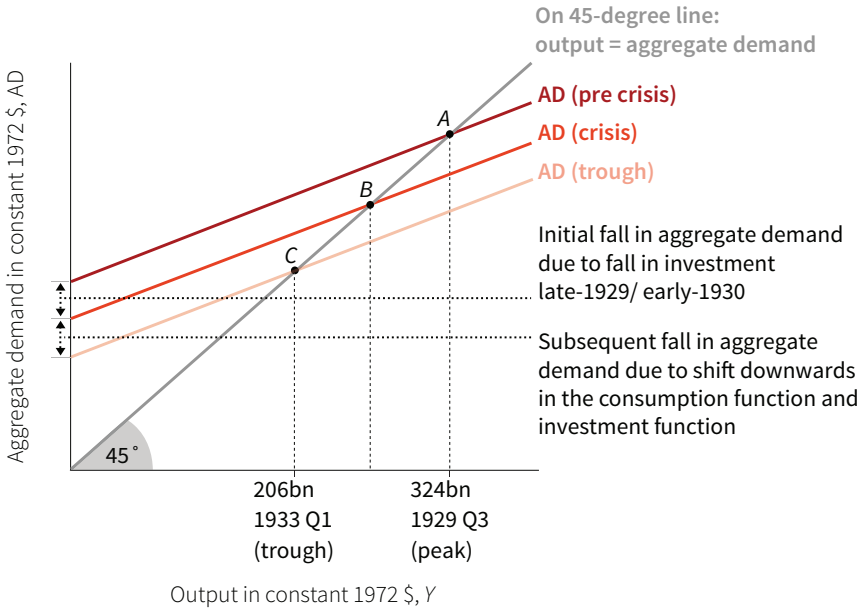


Figure 13.6 *Aggregate demand in the Great Depression: Multiplier and positive feedback processes.*

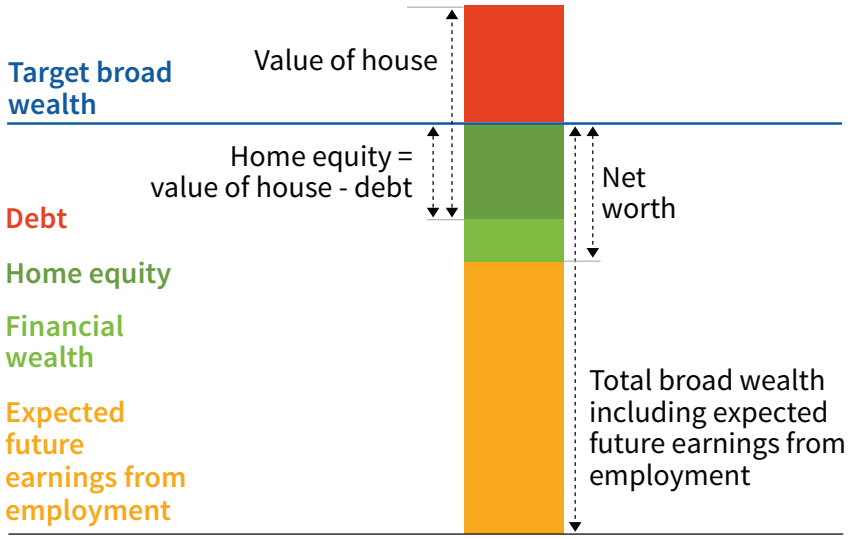
Source: *Gordon, Robert J. 1986. The American Business Cycle: Continuity and Change. Chicago, Il: University of Chicago Press.*

Research done since the Great Depression (which we examine in more depth in Unit 17) provides a number of explanations for the fall in autonomous consumption in the US:

- *Uncertainty:* One channel operated through the effect of the uncertainty about the state of the economy provoked by the dramatic stock market crash of October 1929, which made both firms and households more cautious, prompting them to postpone purchases of machinery and equipment and consumer durables.
- *Pessimism and the desire to save more:* Households also changed their beliefs about their ability to maintain levels of spending as their views about their expected earnings from employment into the future became more pessimistic. Their assessment of their material wealth was also affected as the prices of houses and financial assets fell. The 1920s had seen a build-up of debt by households, as they were able to use instalment agreements for the first time to buy consumer durables.
- *The banking crisis and the collapse of credit:* A third factor that shifted the aggregate demand line down to the level labelled "trough" was the banking crisis of 1930 and 1931, which affected both consumption and investment. There was a wave of failures of small, weak and largely unregulated banks across the US. The system

of small banks was very prone to panic when savers began to fear that they would not be able to get access to their deposits. As explained in Unit 11, as panic spread from bank to bank, bank runs affected the entire banking system. With the collapse of the banking system, households lost deposits and small firms lost their access to credit.

To illustrate why households who were not affected by credit constraints nevertheless cut consumption, we look at the composition of a household's wealth or assets. In Unit 11, we introduced the concept of wealth by comparing it with the volume of water in a bathtub. At that time we focused on material wealth. In Figure 13.7, we extend the concept of wealth to broad wealth so as to include the household's expected future earnings from employment.



For the household shown in the figure, expected broad wealth (orange + light green + dark green) is equal to target wealth.

Figure 13.7 *Household wealth: Key concepts.*

In Figure 13.7, the household's total broad wealth is shown at the top of the dark green rectangle.

Households also hold debt, which is shown by the red rectangle. We have already taken this into account in calculating the household's wealth. Why? Because the household's net worth takes the total of what the household has (its assets) and subtracts the debt it owes. It simplifies matters if we assume that all of the household's debt is the mortgage on the house. This allows us to show that the value of the house (that is, what the household could expect to sell the house for now) is equal to the household's equity in the house plus what it owes to the bank (the mortgage).

As we shall see:

- *If target wealth is above expected wealth:* The household will *increase* savings and decrease consumption.
- *If target wealth is below expected wealth:* The household will *decrease* savings and increase consumption.

DISCUSS 13.1: A HOUSEHOLD'S BALANCE SHEET

Consider a family of two parents and two children who have a mortgage on their home. They have paid off half the mortgage. The family also owns a car and a portfolio of shares in companies. They spend their income on food, clothing and private school fees and have retirement savings held in a pension fund.

1. Which of these items would be on a balance sheet for the household?
2. Think of some typical values for these items in your country for a family of this type. Using the example of the bank's balance sheet in Figure 11.15 as a guide, construct a balance sheet for your hypothetical household.

In early 1929, how would a household with the wealth position shown in *column A* of Figure 13.8 have interpreted news about factory closures, the collapse of the stock market, and bank failures? How would it have adjusted spending on consumer durables, housing and non-durables? Answers to these questions help tell us why the Great Depression happened.

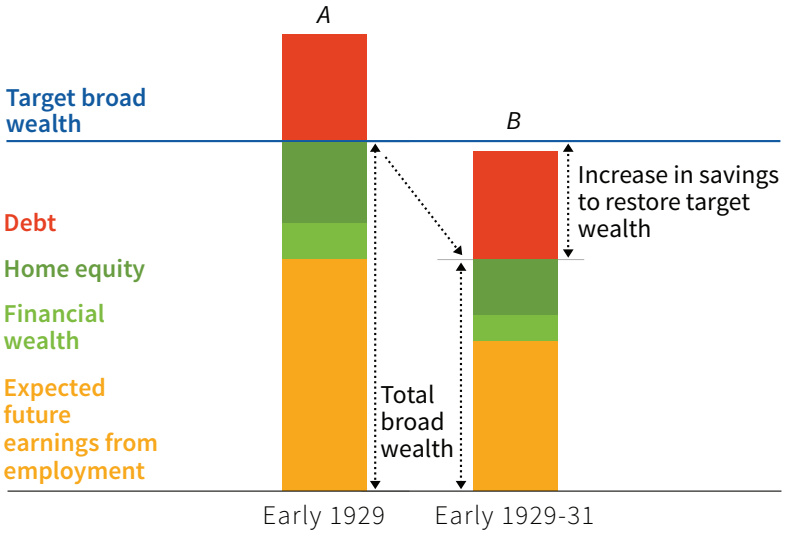


Figure 13.8 *The Great Depression: Households cut consumption to restore their target broad wealth.*

- *Before the Depression:* Viewed from early in 1929 (*column A*), households are shown as making consumption decisions in line with their expectations: total wealth is equal to target wealth.
- *The Depression:* By late 1929 (*column B*), the downturn was underway and beliefs had changed. With job losses throughout the economy, households revised expected earnings downward. Falling asset prices (of shares and houses) reduced the value of the household's material wealth. The result was a gap between the household's target wealth and expected wealth. This helps to explain the cutback in consumption by households who could and—in an ordinary downturn—would have helped smooth a temporary fall in aggregate demand. Instead, these households increased their saving. This fall in autonomous consumption is part of the explanation for the downward shift of the aggregate demand curve from crisis to trough in the multiplier diagram in Figure 13.6.
- *The financial accelerator, collateral and credit constraints:* Changes in household wealth affect consumption through another channel. In Unit 11, we saw that having collateral may enable a household to borrow. An important example is the case of home loans, where the bank extends a loan using the value of the house as collateral. If the borrower fails to keep up the mortgage payments, the bank can repossess the house.

How does an increase in house prices affect consumption?

- *For those who are not credit-constrained:* If the value of your house increases, this improves your net worth and raises your wealth relative to target. We would predict that this would reduce your precautionary savings, increasing consumption.
- *For those who are credit-constrained:* A rise in the price of your house can lead you to increase your consumption spending because the higher collateral enables you to borrow more. The mechanism through which a rise in the value of collateral by relaxing credit constraints leads to an increase in borrowing and spending by households and firms is called the financial accelerator.

DISCUSS 13.2: HOUSING IN FRANCE AND GERMANY

In France and Germany, it is much more difficult for a household to increase its borrowing based on an increase in the market value of the house. In addition, large downpayments (as a percentage of the house price) are required for house purchases.

1. On the basis of this information, how would you expect a rise in house prices in France or Germany to affect spending by households?
2. In the US or UK loans are more easily available based on a rise in home equity and only a small downpayment is required. How would you expect your answer to question 1 to change when considering the US or UK?
3. What do you conclude about the role of the financial accelerator in France and Germany compared with the UK and the US?

This article will tell you more about the influence of a change in house prices on spending in Europe and the US.

13.4 INVESTMENT SPENDING

In Unit 12, we contrasted the volatility of investment with the smoothness of consumption spending. But how do firms make investment decisions? Think of the manager or owner of a firm with some accumulated revenues in excess of costs, deciding what to do with them. There are four choices:

1. *Dividends:* Allocate the funds to managerial or employee salaries or to dividends for owners.
2. *Consumption:* Buy an interest-bearing financial asset such as a bond, or retire (pay off) existing debt.
3. *Investment abroad:* Build new productive capacity in another country.
4. *Investment at home:* Build new capacity in the home country.

The fourth choice is called investment in our model (the third choice is also investment, but it is not spending on the home country's output so it is measured in the foreign country's national accounts as part of their I).

To decide among the four, and assuming that there is no reason to change salaries, the owner with funds now considers the following alternatives, just like Marco in Unit 11:

- *If the funds are disbursed to owners in the form of dividends:* Then, compared with the other options, owners will have more disposable income. So they have a choice of whether to consume more now or later.

- *If the decision is to consume later:* They can either lend (buy a financial asset such as a bond or retire debt) or invest in a new project.
- *If the decision is to invest:* Whether they do it in the home country or abroad will depend on the expected rate of profit for the investment projects under consideration in the two locations.

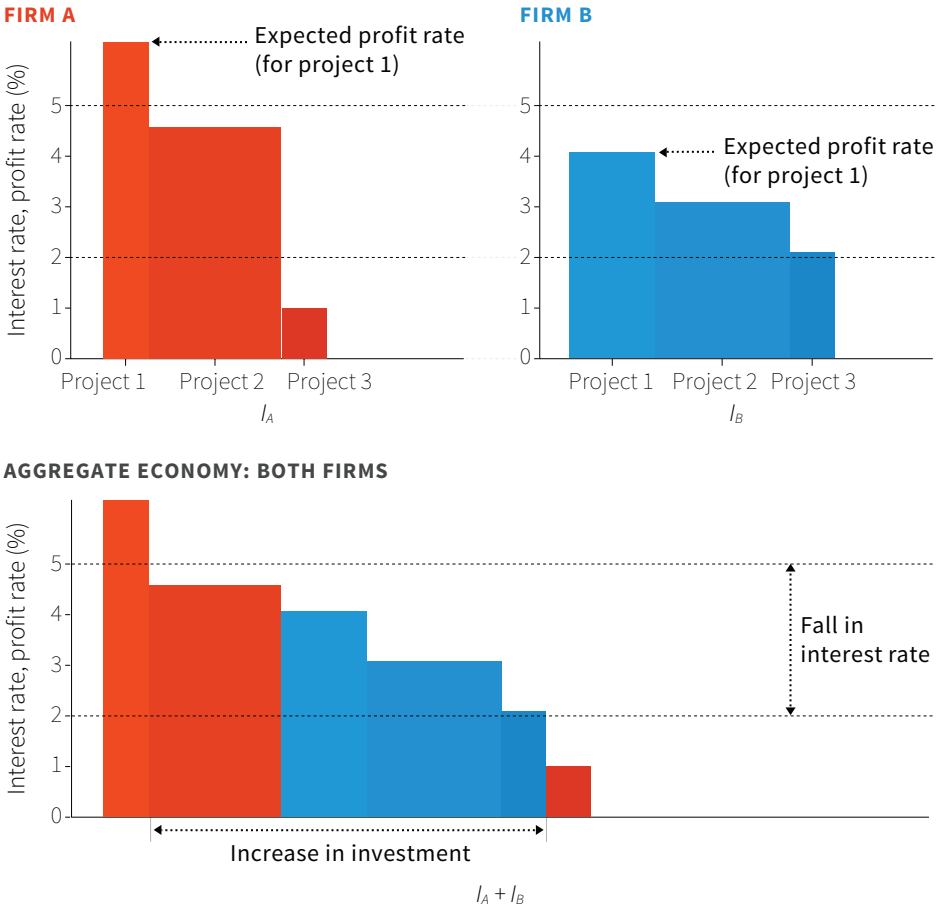
The desirability of consuming more now rather than later depends on the owner's discount rate ρ, as discussed in Unit 12. The owner will compare this to the return they can get by not consuming now: if the firm saves by buying a financial asset then the return is the interest rate r; if the firm invests in productive capacity then the return will be the profit rate on investment, which we will call p:

- *If ρ is greater than both r and p:* The owner will keep the funds and increase consumption spending.
- *If r is greater than ρ and p:* The decision will be to repay debt or purchase a financial asset.
- *If p is greater than ρ and r:* The owner will invest (either at home or abroad).

Because of these options, the interest rate is one of the factors determining whether investment takes place. We saw in Unit 11 that this can be altered by central bank policy. The interest rate is the opportunity cost of purchases of machinery, equipment and structures that increase the capital stock—if you have money available, you could save it with a return of r instead of investing it. Alternatively, if you do not have money available, then the cost of borrowing for investment is also r. If we rank investment projects by their expected post-tax rate of profit, then a lower interest rate raises the number of projects for which the expected rate of profit is greater than the interest rate. We saw this when Marco faced the decision of whether or not to invest (Figure 11.10). Thus a higher interest rate reduces investment, and a lower interest rate increases it.

Figure 13.9 illustrates this for an economy consisting of two firms, *A* and *B*. For each firm in this example, there are three investment projects of different scale and rate of return. They are shown in decreasing order of the expected rate of profit. If the interest rate is 5%, firm *A* goes ahead with project 1 and firm *B* does not invest at all. If the interest rate is 2%, projects 1 and 2 in firm *A* and all three projects in firm *B* are undertaken. The lower panel aggregates the two firms to show how investment in the economy as a whole responds to a change in the interest rate.

In Figure 13.10, we look at how a change in profit expectations affects investment. In the two-firm economy in Figure 13.10a, the expected rate of profit for each project rises because of an improvement in the supply-side conditions in the economy. The height of each column rises and, as a result, there is more investment at a given rate of interest.



Investment in the economy increases after a fall in the interest rate. Five projects go ahead, instead of just one.

Figure 13.9 *Investment, expected rate of profit and interest rate in an economy with two firms.*

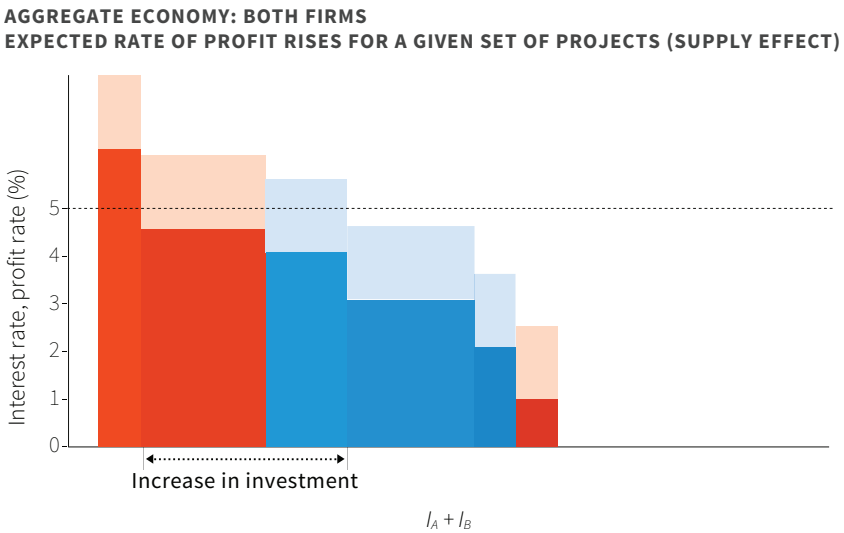


Figure 13.10a *Aggregate economy, where the expected rate of profit rises for a given set of projects (supply effect).*

An upward shift can be caused by a fall in expected input prices, such as a forecast fall in the price of energy or wages, or a fall in taxation over the life of the project. Another example of a positive supply effect is an improvement in the security of property rights so that there is a lesser chance that the government or another powerful actor (such as a powerful landowner, like Bruno in Unit 5, who might threaten a smallholder) will take over ownership of the investment project. This is called a fall in the *risk of expropriation* and is an example of an improvement in the environment for doing business.

In Figure 13.10b, the height of the columns remains unchanged, but the amount of investment that is profitable in many projects has increased. This is the result of a permanent increase in demand and the lack of sufficient capacity to meet forecast sales.

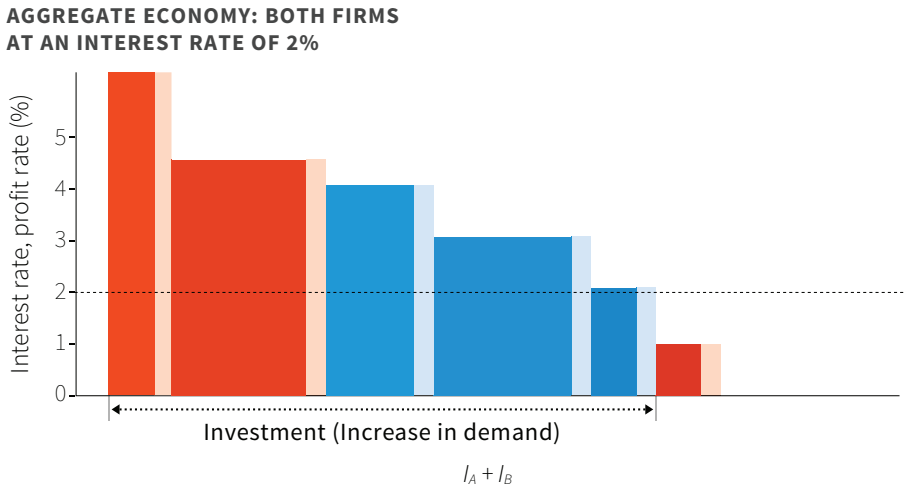


Figure 13.10b *Aggregate economy, where the desired capacity rises for each project (demand effect).*

In an economy with many thousands of firms, a downward-sloping line, as in Figure 13.10c, represents the potential investment projects. This is called the *aggregate investment function*. In addition to the response of investment to a change in the interest rate shown as a shift from *C* to *E*, the figure shows the effect of a change in the profitability of investment, which arises from supply and demand effects and raises investment from *C* to *D* at an unchanged rate of interest.

The empirical evidence suggests that business spending on machinery and equipment is not very sensitive to the interest rate. The limited effect of changes in the interest rate on business investment (illustrated by the steepness of the lines in the figure) highlights the importance of the supply and demand side factors that shift the investment function (Figures 13.10a and 13.10b).

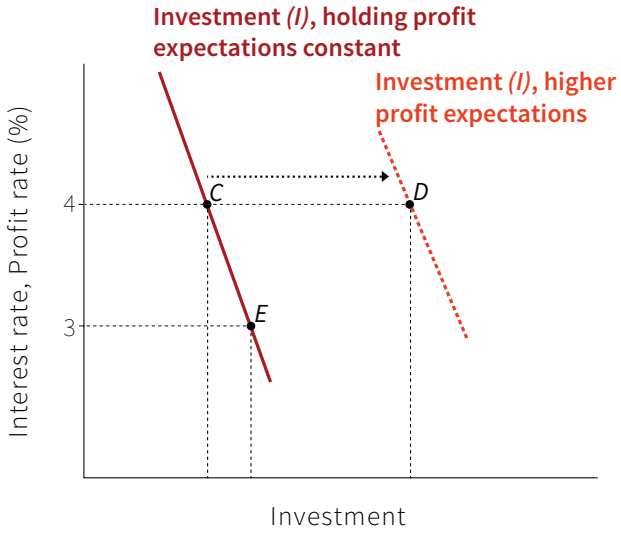


Figure 13.10c *Aggregate investment function: Interest rate and profit expectations effects.*

The interest rate affects investment spending outside the business sector through its effects on households' decisions to purchase new or larger homes, which influence new housing construction. The interest rate also has substantial effects on the demand for durable consumer goods, such as cars and home appliances, which are often purchased using credit.

13.5 THE MULTIPLIER MODEL INCLUDING THE GOVERNMENT AND NET EXPORTS

Here, we add to the model so that we can show how governments and central banks stabilise (or destabilise) the economy after a shock. As before, we assume that firms are willing to supply any amount of goods demanded, such that:

$$\begin{aligned} output &= aggregate\ demand \\ Y &= AD \end{aligned}$$

Adding the government, and interactions with the rest of the world through exports and imports, as we saw in Unit 12, aggregate demand can be split into these components:

$$\begin{aligned} aggregate\ demand &= consumption \\ Y &= AD \\ &\quad + investment \\ &\quad + government\ spending \\ &\quad + net\ exports \end{aligned}$$

To understand the aggregate demand function as shown above, it is useful to go through each component in turn:

Consumption

Household consumption spending depends on post-tax income. The government charges a tax, which we assume is proportional to income at the rate, t. The income left after the payment of tax ($(1 - t)Y$) is called disposable income. The marginal propensity to consume, c_1, is the fraction of disposable income consumed (not pre-tax income). In the aggregate consumption function:

- Spending on consumption is: $C = c_0 + c_1(1 - t)Y$.
- The bigger the fraction of households that are able to smooth fluctuations in their disposable income, the smaller is c_1 and the greater is c_0. Consumption becomes less responsive to income. Therefore the lower the multiplier, and the flatter the consumption function and the aggregate demand curve.
- All of the influences on consumption apart from current disposable income are included in autonomous consumption, c_0, and will therefore shift the consumption function in the multiplier diagram. These include wealth, collateral effects and changes in the interest rate.

Investment

We have just seen that investment spending will be influenced by the interest rate and the expected post-tax rate of profit. In the aggregate investment function:

- Spending on investment is: $I = I(\textit{interest rate, expected post-tax rate of profit})$.
- A higher interest rate, *ceteris paribus,* reduces investment spending, shifting down the aggregate demand curve.
- A higher expected post-tax rate of profit raises investment spending, shifting up the aggregate demand curve.

Government spending

Much of government spending (excluding transfers) is on general public services, health and education. Government spending does not change in a systematic way with changes in income. It is referred to as exogenous.

- An increase in government spending shifts the aggregate demand curve up in the multiplier diagram.

Net exports

The home economy sells goods and services abroad, which are its exports. The amount of goods the home economy demands from abroad, its imports, will depend on incomes at home. The fraction of each additional unit of income that is spent on imports is termed the *marginal propensity to import*, or just m, which must be between 0 and 1. So we have:

$$net\ exports = X - M = X - mY$$

If a country's costs of production fall and as a result it sells its goods at a lower price on world markets compared to the prices of other countries, the demand for its exports will increase, and the demand by home residents for imports will fall. We will see in the next unit that the exchange rate affects the prices of a country's goods on world markets. Growth in world markets increases exports.

Putting together each of the components of aggregate demand we have:

$$AD = C_0 + C_1\,(1-t)Y + I + G + X - mY$$

Both taxes and imports reduce the size of the multiplier. Recall that the multiplier tells us the amount by which an increase in spending (such as a rise in autonomous consumption, investment, government spending or exports) raises GDP in the economy. When we include taxation and imports in the model, the indirect ripple effect of a given rise in spending on GDP is smaller. This is because some of people's income goes straight to the government as taxation, and some is used to buy goods and services produced abroad. Because we assume that the government does not increase its spending when taxes go up, and foreign buyers do not import more of our goods when we import more of theirs, this means that some of an autonomous increase in income does not lead to further indirect increases in income in the domestic economy. Like saving, taxation and imports are referred to as leakages from the circular flow of income. The result is to reduce the indirect effects of an autonomous change in spending on aggregate demand, output and employment.

To summarise:

- A higher marginal propensity to import reduces the size of the multiplier, which makes the aggregate demand curve flatter.
- An increase in exports shifts the aggregate demand curve up in the multiplier diagram.
- An increase in the tax rate reduces the size of the multiplier, which makes the aggregate demand curve flatter.

The second part of our Einstein section shows you how to calculate the size of the multiplier in the model once the tax rate and imports are included. To illustrate, we assume a tax rate of 20%, that is 0.2, and a marginal propensity to import of 0.1. Before we introduced the government, we set the marginal propensity to consume, c_1, at 0.6. If we use these numbers in the formula for the multiplier that we calculate in the Einstein, we get the result that the value of the multiplier is $k = 1.6$, around two-thirds of its value calculated without including taxation and imports. In the next section we look at how economists have estimated the size of the multiplier from data, why they disagree, and why it matters.

DISCUSS 13.3: THE MULTIPLIER MODEL

Consider the multiplier model discussed above.

1. Compare two economies, which differ only in their share of credit-constrained households but are identical otherwise. In which economy is the multiplier larger? Illustrate your answer using a diagram.
2. On the basis of your comparison of the two economies, would you expect the multiplier to vary over the business cycle?
3. Some economists estimated the size of the multiplier in the Great Depression to be equal to 1.8. Explain how the following characteristics of the US economy at the time would have been expected to affect its value: the size of government (see Figure 13.1), the fact that there were no unemployment benefits, and the fact that the share of imports was small.

13.6 FISCAL POLICY: HOW GOVERNMENT CAN DAMPEN AND AMPLIFY FLUCTUATIONS

There are three main ways that government spending and taxation can dampen fluctuations in the economy:

- *The size of government:* Unlike private investment, government spending on consumption and investment is usually stable. Spending on health and education, which are the two largest government budget items in most countries, does not fluctuate with capacity utilisation or move with business confidence. These kinds of government spending stabilise the economy. As we have also seen, a higher tax rate dampens fluctuations because it reduces the size of the multiplier.
- *The government provides unemployment benefits:* Although households save to smooth fluctuations in income, the probability of job loss is low for an individual so that person will not save enough (that is, self-insure) to cope with an extended period of unemployment. Other programs to redistribute income to the poor have the same smoothing effect.
- *The government can intervene:* It can intervene deliberately to stabilise aggregate demand using fiscal policy.

Could workers insure privately against job loss? There are also three reasons why the private market fails, and therefore governments provide unemployment insurance:

- *Correlated risk:* In a recession, job loss will be widespread. This means that there will be a surge in insurance claims across the economy and a private provider may be unable to pay out on the scale required. It also means co-insurance among a group of neighbours or family members may be of limited use as the need for help may arise in many households at the same time.
- *Hidden actions:* As we saw in Unit 10 the insurance company cannot observe the reason for the job loss so it would have to insure the employee not only against a firm cutting back employment due to lack of demand but also against the worker being fired for inadequate work. This creates a *moral hazard*, because a well-insured person would be expected to make less of an effort on the job. By promoting exactly the behaviour that the company is insuring against, this hidden actions problem reduces the profits of the insurance company.
- *Hidden attributes:* Suppose you learn that your firm is in difficulty, but the insurance company does not. This is another example of *asymmetric information*. You will therefore buy insurance when you learn of the likely closure of the firm and it will be provided at reasonable rates because the insurance company does not know that you are likely to make a claim on them. Workers who know their firm is performing well will not buy insurance. The hidden attributes problem will be true about individuals (hardworking or lazy), not just firms (successful or failing). The good prospects (those who enjoy working hard, for example) will shun the insurance and the insurer will be left with those likely to take the extra risks of losing their job.

Government support to household income during spells of unemployment dampens the business cycle. As we saw in Unit 6, when someone loses his or her job, that person loses wage income. The unemployment insurance or unemployment benefit system provides replacement income. The unemployment insurance payments can normally be claimed for a given number of weeks, and they are much lower than a wage. This difference is the employment rent received by a worker with a job.

The system of unemployment benefits is part of the *automatic stabilisation* that characterises modern economies. We have already seen one automatic stabiliser: a proportional tax system reduces the size of the multiplier and dampens the cycle.

AUTOMATIC STABILISERS

Characteristics of the tax and transfer system in an economy that have the effect of offsetting an expansion or contraction of the economy.

In our list, the third role of government in dampening fluctuations is the use of fiscal policy in deliberate stabilisation policies: upward adjustment of spending, or cuts in taxation, to support aggregate demand in a downturn; or trimming spending and raising taxes to rein in a boom. But it is cumbersome to have these fiscal policy

measures approved by a parliament, which has power over budgetary decisions—one reason that stabilisation policy is often handled through monetary, rather than fiscal, policy.

How government can amplify fluctuations

A government might choose to override the automatic stabilisers because it is concerned about the effect of a recession on its *budget balance*. Government budget balance is the difference between government revenue less transfers, T, and government spending, *G*, that is, *(T – G)*. As we have seen, if the economy is in recession, government transfers rise while tax revenues fall, so the government's budget balance deteriorates and may become negative.

When the government's budget balance is negative, this is called a *government budget deficit*—government spending on goods and services, including investment spending, plus spending on transfers (such as pensions and unemployment benefits) is greater than government tax revenue. A government budget surplus is when tax revenue is greater than government spending. To summarise:

- *Budget in balance:* $G = T$
- *Budget deficit:* $G > T$
- *Budget surplus:* $G < T$

The worsening of the government's budgetary position in a recession is part of its stabilising role. When it chooses to override the stabilisers to reduce its deficit, this may amplify fluctuations in the economy.

The fallacy of composition and the paradox of thrift

By comparing a household with the economy as a whole, we understand better the nature of an increase in the government's deficit in a recession. Faced with a household budget deficit, a family worried about mounting debts as their wealth falls further below their target cuts spending and saves more. We saw exactly this behaviour in Figure 13.8 when households increased their savings in 1929. Keynes showed that the wisdom of family precautionary saving does not apply to the government when the economy is in a recession.

Compare the attempt to save more by a single household and by all households in the economy simultaneously. Think of a single household putting its additional savings in a sock. The money is in the sock for when the household decides it is wise to spend it.

Now, assume that all households put additional savings in their socks. Assuming nothing else in the economy changes, the additional saving causes lower aggregate consumption spending in the economy. What happens? From the previous section, we can model this as a fall in autonomous consumption, c_0: the aggregate demand curve shifts down. The economy moves through the multiplier process to a lower

level of output, income and employment. As incomes fall, so do savings: the families take money out of their socks and spend it. Once the economy is at the new lower output and employment equilibrium, there is no money left in the socks.

A single household can increase its saving if it anticipates bad luck, and the saving will be there if it is unlucky—for example, if someone becomes ill or loses a job. However, if every household does this when the economy is in a recession, this behaviour causes the bad luck: more people lose their jobs. The reason: in the economy as a whole, spending and earning go together. My spending is your income; your spending is my income.

What can be done? The government can allow the automatic stabilisers to operate and help absorb the shock. In addition, it can provide an economic stimulus (such as a temporary increase in government spending or a temporary cut in taxation) until business and consumer confidence return and the private sector regains its willingness to spend. Budget deficits rise, but this avoids a deep recession, as Keynes realised.

GREAT ECONOMISTS

JOHN MAYNARD KEYNES

John Maynard Keynes (1883-1946), and the Great Depression of the 1930s, changed the course of economics. Until then most economists had seen unemployment as the result of some kind of imperfection in the labour market. If this market worked optimally it would equate the supply of, and demand for, workers. The massive and persistent unemployment in the decade prior to the second world war led Keynes to look again at the problem of joblessness.

Keynes was born into an academic family in Cambridge, UK. He studied mathematics at King's College, Cambridge and then became an economist and prominent follower of the renowned Cambridge professor, Alfred Marshall. Before the first world war Keynes became a world authority on the quantity theory of money and the gold standard, and held conservative views on economic policy, arguing for a limited role of government. But his views would soon change.

In 1919, following the end of the first world war, Keynes published *The Economic Consequences of the Peace*, which opposed the Versailles settlement that ended the war. This book instantly made him a global celebrity. Keynes rightly argued that Germany could not pay large reparations for the war, and that an attempt to make Germany do this would help provoke a worldwide economic crisis.

In 1925 Keynes opposed Britain's return to the gold standard, arguing that this policy would lead to a contraction of the economy. In 1929 there was a financial crash and global crisis, and in 1931 Britain was driven off the gold standard. The Great Depression followed.

In response to these dramatic events, Keynes explained that the orthodox monetary policies required by the gold standard would worsen the depression, and that the world needed policies to increase aggregate demand. In 1936, he published *The General Theory of Employment, Interest and Money* in which he set out an economic model to explain these views. The General Theory immediately became world famous, particularly for the idea of the multiplier, which is explained in this unit. In the General Theory, Keynes reasoned that if interest rates were already very low, then fiscal expansion would be necessary to alleviate depression. The initial response in many countries to the global economic crisis of 2008 was to apply such Keynesian policies.

During the second world war, Keynes turned to postwar reconstruction, determined to ensure that the mistakes that followed the first world war would not be repeated. In 1944, with Harry Dexter White of the US, he led an international conference at Bretton Woods in New Hampshire that resulted in the creation of a new international monetary system, managed by the International Monetary Fund, or IMF. The Bretton Woods system was designed to avoid the mistakes Keynes had unsuccessfully warned against in the aftermath of the first world war, and to ensure that a country that was in recession (and had balance of payments difficulties) would not need to follow the contractionary policies required by the gold standard. A country like this could use fiscal policy to pursue full employment; at the same time it could devalue its exchange rate to encourage exports, reduce imports, and achieve a satisfactory balance of payments position.

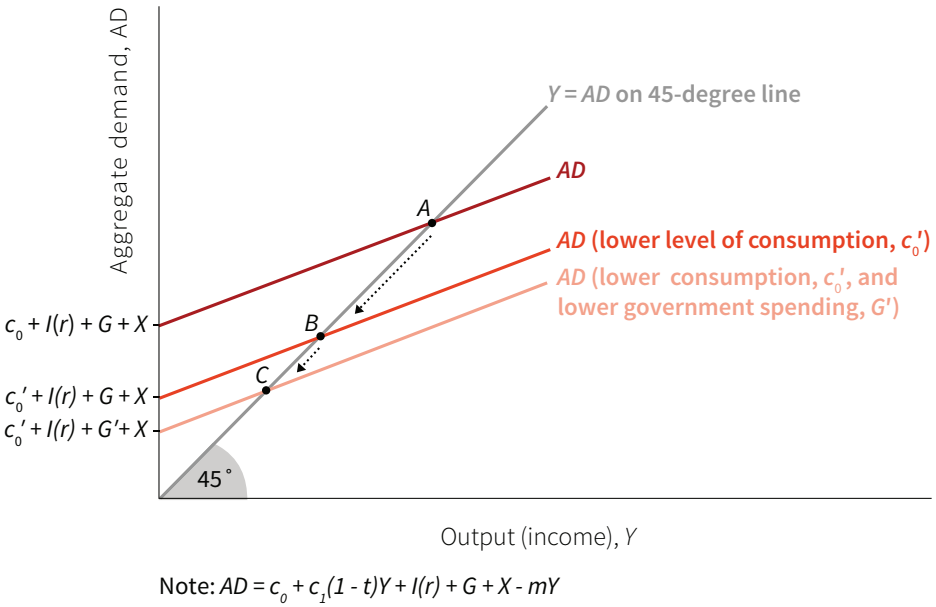
Keynes led a remarkably varied life. He was an academic, a senior civil servant, owner of the New Statesman magazine, financial speculator, chairman of an insurance company, and member of the British House of Lords. He was also the founder of the Arts Council of Great Britain and chairman of the Covent Garden Opera Company. He was married to the Russian ballerina Lydia Lopokova, and was a key member of the Bloomsbury Group, a remarkable circle of artistic and literary friends in London, which included the novelist Virginia Woolf.

In 1926 in a pamphlet entitled The End of Laissez-Faire he wrote:

"For my part I think that capitalism, wisely managed, can probably be made more efficient for attaining economic ends than any alternative system yet in sight, but that in itself it is in many ways extremely objectionable. Our problem is to work out a social organisation which shall be as efficient as possible without offending our notions of a satisfactory way of life."
– John Maynard Keynes, *"The End of Laissez Faire" (1926)*

Keynes' argument refers to the cell in the bottom right of Figure 13.12 at the end of this section: poor policymaking that amplifies the business cycle. Suppose a government tries to improve its budgetary position in a recession by increasing its saving. This is referred to as *austerity policy*.

Figure 13.11 shows in the multiplier diagram how austerity policy can reinforce a recession. The economy starts at point A in goods market equilibrium, at which aggregate demand is equal to output. The economy then moves into recession after a fall in consumer confidence, which reduces c_0, the part of consumption that does not depend on income, to c_0'. The aggregate demand line shifts downward and the economy moves from point *A* to point *B*. Lower output reduces tax revenues to the government and increases spending on unemployment benefits, worsening the budget balance. Suppose that the government then reduces spending, from *G* to *G'*, in a bid to offset this deterioration of its budget balance. But this pushes the aggregate demand line down further and the economy moves to point *C*; the government action worsens the recession. The recession feeds back to raise government transfers and reduce tax revenue.



Note: $AD = c_0 + c_1(1 - t)Y + I(r) + G + X - mY$

Figure 13.11 *Government austerity can worsen a recession.*

DISCUSS 13.4: SPENDING CUTS IN A RECESSION

Assume the government is initially in budget balance.

Does the government's budget balance improve, deteriorate or remain unchanged if the government cuts its spending in a recession, ceteris paribus? Evaluate the government's policy.

Does this argument mean that government deficits should never be a concern? No. Running government deficits under the wrong economic conditions can be harmful. In a well-designed policy framework, there will be constraints on government action, as we will see in section 13.8.

Figure 13.12 summarises the lessons so far. The first row gives examples of how household behaviour may either smooth or disrupt the economy. The terms *negative* and *positive feedback* are used to refer to dampening and amplifying mechanisms in the business cycle.

POSITIVE AND NEGATIVE FEEDBACK PROCESSES

- *Positive feedback:* A process whereby some initial change sets in motion a process that magnifies the initial change.
- *Negative feedback:* A process whereby some initial change sets in motion a process that dampens the initial change.

	DAMPENING MECHANISMS OFFSET SHOCKS (STABILISING)	AMPLIFYING MECHANISMS REINFORCE SHOCKS (MAY BE DESTABILISING)
PRIVATE SECTOR DECISIONS	Consumption smoothing	Credit constraints limit consumption smoothing Rising value of collateral (house prices) can increase wealth above the target level and raise consumption Rising capacity utilisation in a boom encourages investment spending, adding to the boom
GOVERNMENT AND CENTRAL BANK DECISIONS	Automatic stabilisers (e.g. unemployment benefit) Stabilisation policy (fiscal or monetary)	Policy mistakes, such as limiting the scope of automatic stabilisers in a recession or running deficits during low demand periods, while not running surpluses during booms

Figure 13.12 *The role of the private sector and the government in the business cycle.*

13.7 THE MULTIPLIER AND ECONOMIC POLICYMAKING

In this unit we examine the role of government spending or tax cuts to stabilise the economy. In the next unit we focus on the use of monetary policy to shift the aggregate demand curve following a shock.

In the multiplier model, we have used simple ways of modelling aggregate consumption, investment, trade and government fiscal policy. This means there are a small number of variables from which the size of the multiplier is calculated (the marginal propensity to consume, the marginal propensity to import, and the tax rate). When we move to apply the model to the world, it is important to realise that there is no single multiplier that applies at all times.

The multiplier will be a different size if the economy is operating at full capacity utilisation and low unemployment than in a recession: with fully employed resources, a 1% increase in government spending would displace or *crowd out* some private spending in the economy. To consider an extreme case, if all workers are employed, then an increase in government employment can only come about by taking workers out of the private sector.

In these conditions we would not expect a positive multiplier, let alone a multiplier greater than one. Moreover, we would not expect a government to decide to undertake a fiscal expansion in a fully employed economy (although it may have to do this if the country is at war, as the US did in the later years of the second world war and in the Vietnam war).

The size of the multiplier will also depend on the expectations of firms and businesses. The economy is not like a bicycle tyre, from which air can be pumped in or let out to keep the pressure at the right level. Households and firms react to policy changes, but they also anticipate them. For example, if firms anticipate that the government will stabilise following a negative shock, this will support business confidence, and the policymaker will be able to use a smaller stimulus. Alternatively, if households think that higher government spending will be followed by higher taxes, those who have the ability to save may put aside more of their money to pay the extra taxes. If this happened, it would reduce the impact of the stimulus.

When the financial crisis in 2008 led to the biggest falls in GDP in many economies since the Great Depression, the world's policymakers expected an answer from economists: would fiscal policy help to stabilise the economy? The multiplier model, inspired by Keynes' analysis of the Great Depression, suggested that it would. But by 2008 many economists doubted that the Keynesian model was still relevant. The crisis has revived interest in it and has led to a greater, though not complete, consensus among economists about the size of the multiplier (see below).

In 2015 a study published by Alan Auerbach and Yuriy Gorodnichenko, two economists, shows how the multiplier varies in size according to whether the economy is in a recession or in an expansion. This is exactly the question policymakers needed an answer to in 2008.

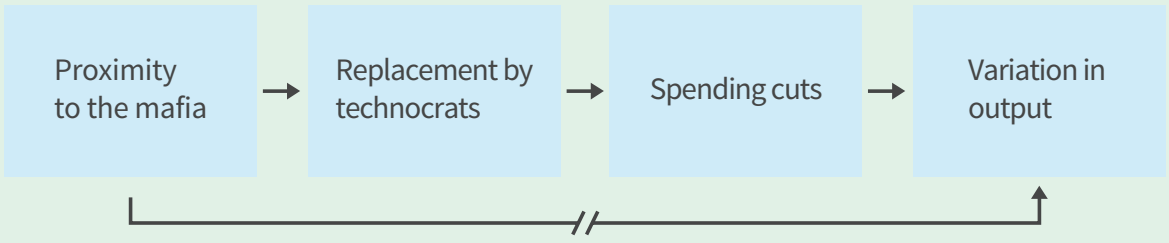
For the US, it suggested a $1 increase in government spending in the US raises output by about $1.50 to $2.00 in a recession, but only about $0.50 in an expansion. Auerbach and Gorodnichenko extended their research to other countries and found similar results. They also found that the effect of autonomous increases in spending in one country had spillover effects on the countries with which they trade. These effects were about the same magnitude as the indirect effects of second, third, and further rounds of spending in the home country.

HOW ECONOMISTS LEARN FROM FACTS

THE MAFIA AND THE MULTIPLIER

It may surprise you that economists have used the Italian government's struggle against the mafia to uncover the size of the multiplier, but that's what Antonio Acconcia, Antonio, Giancarlo Corsetti, and Saverio Simonelli were able to do. Adopting the natural experiment method to address the problem of reverse causality, they used data on mafia-related dismissals of local politicians to isolate a variation in public spending that is not caused by variation in output.

After a legal change in 1982, if provincial councils in Italy were revealed to have close links with the mafia the central government dismissed them, and appointed new officials in their place. These technocrats cut local spending by 20% on average. The change in public spending occurred because of the mafia links through their effect on the replacement of government officials. And because there is no direct causal link from proximity to the mafia to the variation in output, proximity to the mafia can be used to uncover the causal effect of a change in public spending on output. This is illustrated below:

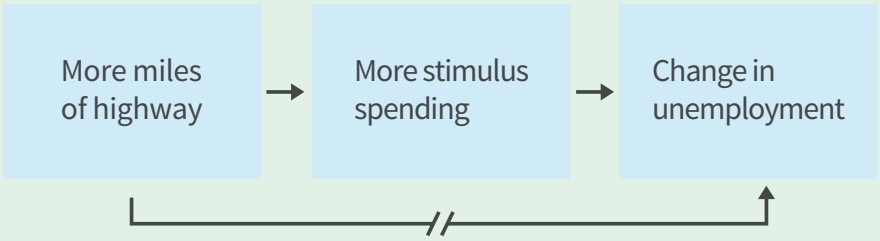


Using this method, the researchers were able to estimate multipliers of 1.5 at the local level.

Economists have used their ingenuity to come up with methods of estimating the size of the multiplier and the implication of its operation for jobs. Using the US stimulus programme that was implemented in the wake of the financial crisis (the American Recovery and Reinvestment Act of 2009, a $787bn fiscal stimulus) we would expect that US states that were more severely hit by the financial crisis would have had more unemployment and attracted more stimulus spending by the government. So the unemployment causes more spending in those states. This makes it difficult to estimate the effect of higher spending on output and unemployment, which is what we want to do if we want to know the size of the multiplier.

One approach to get around this problem of reverse causality is to make use of the fact some of the spending in the US stimulus programme was distributed to US states using a formula that was unrelated to the severity of the recession experienced in each state. For example, some road-repair expenditures funded by the stimulus package were based on the length of highway in each state.

Given the formula for allocating road-building funds and the fact that more miles of highway has no direct effect on the change in unemployment, this allows us to answer the question: were more jobs created in states that received more stimulus spending?



The results of studies using this approach estimated the multiplier equals two, and suggest that the Act created between 1 million and 3 million new jobs.

In spite of scepticism among some economists before the 2008 crisis that the multiplier was greater than one, policymakers around the world embarked on fiscal stimulus programmes in 2008-09. Fiscal stimulus was credited with helping to avert another Great Depression, as we will see in Unit 17.

WHEN ECONOMISTS DISAGREE

HOW RESPONSIVE IS THE ECONOMY TO GOVERNMENT SPENDING?

There are few questions in economic policy discussed as heatedly in the years since the financial crisis in 2008 as the size of the fiscal multiplier: what is the effect on GDP of a 1% increase in government spending?

Much of the heat is generated by political differences among those involved: people who favour greater government expenditure tend to think the multiplier is large, while those who would like to shrink government tend to think that it is small. (We don't know whether this correlation is because their beliefs about government influence their estimates of the size of the multiplier, or the other way round.)

This debate has been going on since the first theoretical formalisation of the multiplier by John Maynard Keynes in the 1930s. The recent economic crisis has revitalised it for two reasons:

Monetary policy could not be used: Following the financial crisis, GDP fell and governments wanted to know if a fiscal stimulus of an increase in government spending would help stabilise the economy. The focus was on fiscal stimulus because monetary policy had already been used to cut interest rates all the way down to the zero lower bound.

Does austerity work? After the eurozone crisis in 2010, many countries in Europe, which were in recession, adopted the opposite fiscal policy to stimulus. They adopted austerity measures of cutting government spending with the objective of bringing their public finances back to balance.

In both stimulus and austerity, the success of the policy depends on the size of the multiplier. If the multiplier is negative, a stimulus package would lead to a reduction in GDP, and an austerity policy would cause GDP to rise. If the multiplier is positive but less than one, a fiscal stimulus would raise GDP but by less than the increase in government spending. If, as in our multiplier model, the multiplier is greater than one, a fiscal stimulus would raise GDP by more than the increase in government spending and a policy of austerity would reinforce the recession conditions.

Depending on methodologies and assumptions, economists have put forward different estimates of multipliers, from negative numbers to values greater than two. For instance, members of President Obama's Council of Economic Advisors estimated the multiplier as 1.6 when they prepared the American Recovery and Reinvestment Act of 2009. The International Monetary Fund presented estimates in 2012 that multipliers in advanced economies were, after the crisis, between 0.9 and 1.7.

To be effective, government spending needs to put resources that would otherwise be idle into productive use. These resources can be unemployed (or underemployed) workers as well as offices, shops or factories functioning with spare capacity. When an economy functions at full capacity (with no idle resources) extra government spending will crowd out private spending.

Economists Robert Barro and Paul Krugman disagreed about the size of the multiplier in the weeks that followed the enactment of the American Recovery and Reinvestment Act in early 2009. Using data on government defence spending during the second world war, Robert Barro concluded that the multiplier was not larger than 0.8. That is, spending $1 on military equipment yielded only 80 cents of output. However, Paul Krugman responded that in wartime there are no idle productive resources to take advantage of. People of working age were in work supporting the war effort in factories, and the government used rationing to depress private consumption.

In the recessions that followed the eurozone crisis in 2010, just as new economic research was finding evidence that multipliers in recessions were well above one, many European governments implemented fiscal austerity to balance their budgets. In another sign that, in deep recessions, the multiplier is greater than one, they had poor growth outcomes. But some eurozone countries had no choice but to adopt austerity policies: as we will see in the next section, they had lost the ability to borrow.

DISCUSS 13.5: METHODS TO ESTIMATE THE MULTIPLIER

Consider the three methods discussed in this unit that have been used to estimate the size of the multiplier: the highways spending in the great recession in the US, the mafia-related dismissals in Italy, and wartime defence spending in the US. Why do you think estimates of the size of the multiplier vary?

Use the material in this unit to support your explanation.

Now you have enough information to see how the contributions to the growth of GDP are divided among consumption, investment, government spending, and net exports.

DISCUSS 13.6: CONTRIBUTIONS TO CHANGE IN REAL GROSS DOMESTIC PRODUCT OVER THE BUSINESS CYCLE

In Figure 12.9 we showed the contributions of the main components of expenditure (*C, I, G* and *X – M*) made to US GDP growth during the recession of 2009. We can use FRED to see whether these contributions changed during the recovery phase of the recession.

- Go to the FRED website
- Search for "Contribution to GDP" using the search bar, and select the annual series:

Contribution to percent change in real gross domestic product: Personal consumption expenditures
Contribution to percent change in real gross domestic product: Gross private domestic investment
Contribution to percent change in real gross domestic product: Government consumption expenditure and gross investment
Contributions to percent change in real gross domestic product: Net exports of goods and services

- Click the "Add to Graph" button to create a graph of the four series.
- Use the "Add Data Series" option to add a series for the growth of real GDP.

1. Do the contributions to GDP add up approximately to the growth of GDP?

Now use the data you have downloaded to carry out the following tasks for the period from 2007 to 2014:

2. Describe the contributions to US GDP growth in the recession (2008Q1 to 2009Q2) and in the recovery phase from 2009Q3 of the business cycle. If you analyse the data using the FRED Graph, you will see the recession shaded in the chart.
3. What might explain the differences seen in the role of consumption and investment during the recession and recovery phases of the business cycle?
4. From the contribution to GDP growth of government consumption and investment expenditure, what can you infer about the US government's fiscal policy during the crisis?

Note: To make sure you understand how these FRED graphs are created, you may want to extract the data in Microsoft Excel format and reproduce the series.

DISCUSS 13.7: THE FALL OF FRANCE

In an article from August 2014 called *The Fall of France*, Paul Krugman criticises the austerity policy implemented in France.

Use what you have learned about the fiscal multiplier to explain why, in Krugman's opinion, fiscal austerity in France (and more generally in Europe) would fail (explain carefully what you think Krugman means by "fail").

DISCUSS 13.8: STIMULUS WITHOUT MORE DEBT

Read this article by Robert Shiller.

Assume the economy is in a recession. The government has a high level of debt and wants to set a balanced budget, G = T. How can the government achieve a fiscal stimulus effect on GDP whilst keeping the budget balanced?

To answer the question, take the following steps.

1. Show how this is possible in a multiplier diagram ensuring that you label the relevant intercepts and angles. Make the diagram sufficiently accurate that the exact size of the multiplier is visible.
2. Explain in words how the government can achieve such a fiscal stimulus effect whilst keeping the budget balanced.
3. Derive the balanced budget multiplier using algebra. (Hint: You will need to write down expressions for the change in GDP associated with a change in both G and T and set these equal.)
4. Comment briefly on any disadvantages you see with the use of this balanced budget fiscal stimulus.

To answer the question, make the following assumptions:

- Assume a lump sum tax. This means that the tax does not depend on the level of income, $T = T$, rather than our usual assumption that $T = tY$.
- Also assume that the country does not have any imports or exports.

13.8 THE GOVERNMENT'S FINANCES

From the paradox of thrift, we learned that in a recession, it is counterproductive for the government to offset the automatic stabilisation of the economy. We have also learned that using a fiscal stimulus to boost aggregate demand, in a deep recession, under conditions in which the multiplier is greater than one, can be justified. So why are stimulus policies often followed by policies of austerity? The answer: the government's debt. To understand why, we turn to the government's revenue and its expenditure.

Revenue

The government raises taxes in the form of income tax and a tax on spending, which in many countries is called Value Added Tax (VAT) or sales tax. It also raises money from taxes on products like alcohol, tobacco, and petrol—and on wealth, including through inheritance taxes.

Expenditure

The government spends on its responsibilities, such as health, education or defence. Government revenue is also used to fund social security transfers, which include unemployment benefits, pensions and disability benefits. The government also has to pay interest on its debt.

Government primary deficit

The government deficit, *excluding interest payments on its debt*, is measured by $G - T$ (or $G - tY$ in the multiplier model with a proportional tax rate, t) and is called the *primary budget deficit*. If the initial situation is one of a zero primary deficit, then it automatically worsens in a business cycle downturn. When the downturn reverses, the government's primary budget deficit will decline and, in the upswing, the government will have higher revenues than spending.

When there is a budget deficit, this means the government must borrow to cover the gap between its revenue and its expenditure. The government borrows by selling bonds. Firms and households buy the bonds. Households usually buy them indirectly, because they are bought by pension funds, from which households buy pensions. The sale of bonds adds to the government's debt.

GOVERNMENT DEBT

The government's debt is:

- The sum of all the bonds it has sold over the years to finance its deficit...
- ... minus the ones that have matured

Because of the existence of global financial markets, foreigners can also buy home country bonds. Government bonds are attractive to investors because they pay a fixed interest rate. They are generally considered a safe investment because the government is expected to pay the interest, and to pay back the principal when it is due. This is another way of saying that the default risk on government bonds is usually low. Investors are likely to want to hold a mixture of safe and risky assets, and government bonds are normally at the safe end of the spectrum.

A sovereign debt crisis is a situation in which government bonds come to be considered risky. In 2010, there was an increase in interest rates on bonds issued by the Irish, Greek, Spanish and Portuguese governments, which was a signal of a sharp increase in default risk. It marked the start of the eurozone crisis. Governments of countries in a sovereign debt crisis may have no alternative to austerity policies if they can no longer borrow, because in this case they cannot spend more than the tax revenue they receive.

A large stock of debt relative to GDP can be a problem because, like a household, the government has to pay interest on its debt and it has to raise revenue to pay the interest. However, governments are not like households in that there is no point at which they need to have paid off all their stock of debt—as one set of bonds matures, governments will typically issue more bonds, maintaining a stock of debt (this is called *rolling over debt*, which firms also typically do to finance their operations). Indeed, because government bonds are generally seen as a safe asset outside periods of crisis, there is usually demand for *government debt* from private investors. As the long-run data for the UK in Figure 13.13 makes clear, there are no general rules about how much debt it is safe for governments to have.

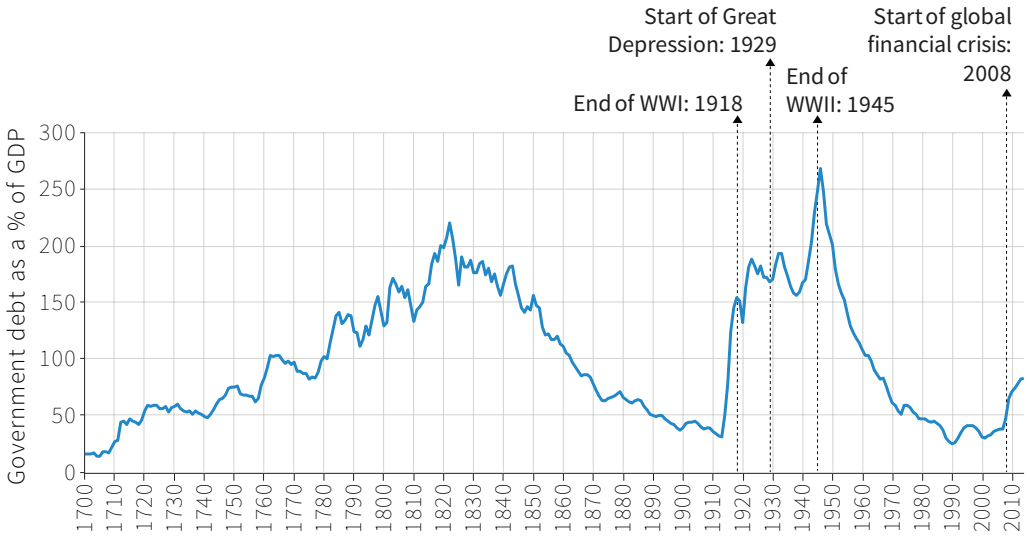


Figure 13.13 *UK government debt as a percentage of GDP, 1700-2014.*

Source: *Bank of England. 2015. 'Three Centuries of Macroeconomic Data.'*

Figure 13.13 shows the path of UK government debt from 1700 to 2014. The level of indebtedness of a government is measured in relation to the size of the economy, that is, as a percentage of GDP. The two big upward spikes in the British debt to GDP ratio in the 20th century were caused by the need for the government to borrow to finance the war effort.

Financial crises also raise government debt: governments borrow both to bail out failing banks and to support the economy in the lengthy recessions that follow financial crises. The UK's debt-to-GDP ratio rapidly doubled to more than 80% after the global financial crisis in 2009.

Note also that, although the UK government emerged from the second world war with a very high level of debt, it fell rapidly during the 1950s and 1960s: from 260% of GDP to 50% by the 1980s. Why? The British government ran a primary budget surplus in every year except one from 1948 until 1973, which helped to reduce the ratio of debt to GDP. But it may also fall even when there is a primary budget deficit, as long as the growth rate of the economy is higher than the interest rate. During the period of rapid reduction of the British debt ratio, in addition to the primary surpluses, there was moderate growth, low nominal interest rates set by the government, and moderate inflation.

Why does inflation help a country reduce its debt ratio? Because the face value of government bonds (the level of debt) is denominated in nominal terms. For instance, a 10-year bond issued in 1950 would promise to repay £1m in 1960. So if inflation was moderately high during the 1950s, then nominal GDP would be growing fast while that £1m owed in 1960 would remain constant, meaning the debt would have shrunk relative to GDP. As we discuss further in Unit 14, inflation reduces the real value of debt.

For many advanced economies, there have been extended periods in which the growth rate has been higher than the interest rate. Brad de Long, an economist, has pointed out that this has been true for the US for almost all of the last 125 years.

DISCUSS 13.9: EFFICIENCY AND FAIRNESS

How would you use the criteria of *efficiency* and *fairness* to evaluate the use of stimulus policies following the global financial crisis of 2007-2008? Hint: you might want to look back at this section and this section in Unit 5, where the concepts are explained.

Countries with ageing populations have demographic trends that imply upward pressure on the debt ratio, because the proportion of government revenue spent on state pensions, healthcare, and social care for the elderly will increase. Many governments and voters are facing a difficult choice. Do they limit benefits, or put up taxes?

The lessons from our discussion of fiscal policy and government debt:

- *Automatic stabilisers play a useful role:* Over the business cycle, they contribute to economic wellbeing.
- *If additional fiscal stimulus is used, this needs to be reversed later:* This can take place when the economy is growing again. If it is not reversed, the debt ratio will rise.
- *Financial crises and wars increase government debt.*
- *Inflation reduces the debt burden of the government:* Likewise, deflation increases it.
- *An ever-increasing debt ratio is unsustainable.*
- *If the growth rate is below the interest rate, it is necessary to run primary government surpluses:* They stabilise and reduce the debt ratio. Attempting to reduce the debt ratio rapidly, however, is counterproductive if it depresses growth.

To get a feel for the effects of policy interventions, use this tool to experiment as a policymaker: try different combinations of primary budget balance, growth rate, nominal interest rate and inflation rate as methods of preventing the debt ratio from continuously rising in a country of your choice.

13.9 FISCAL POLICY AND THE REST OF THE WORLD

In Unit 12 we saw that agrarian economies suffered shocks from wars, disease and the weather. In Unit 9, we saw that the American Civil War affected economies including Brazil, India and the UK. In modern economies, what happens in the rest of the world is a source of shocks, and also affects how domestic economic policy works. To avoid making mistakes, policymakers need to know about these interactions.

Foreign markets matter

Fluctuations in growth in important markets abroad can explain why the economy moves into an upswing or downswing: this is a change in the net export component of aggregate demand, that is, (X – M). China, for example, is a very important market for Australian exports (32% of Australian exports went to China in 2013, accounting for 6.5% of Australian aggregate demand). When the Chinese economy slowed down from a growth rate of 10.4% in 2010 to 7.7% in 2013, this was transmitted directly to a slowdown in growth in Australia via a fall in its net exports.

Similarly, the slowdown in the eurozone because of the 2010 crisis, which followed the global financial crisis, was an important reason for the sluggishness of the British economy's exit from recession. This is because a high proportion of UK exports go to the EU. For example, 43% of the UK's exports went to the EU in 2013, accounting for 13.5% of UK aggregate demand.

Imports dampen domestic fluctuations

As we have seen, the size of the multiplier is reduced by the marginal propensity to import: when autonomous demand goes up, it stimulates spending and some of the products bought are produced abroad. This dampens the domestic upswing.

Trade constrains the use of fiscal stimulus

Trade with other countries constrains the ability of domestic fiscal policymakers to use stimulus policies in a recession. A striking example comes from France in the 1980s. At the start of the 1980s, the French economy remained weak following the oil shocks of the 1970s, which disrupted the world economy. In 1981, the socialist candidate François Mitterrand won the presidential election. His appointed prime minister, Pierre Mauroy, implemented a programme to stimulate aggregate demand through increased government spending and tax cuts (in the multiplier model, this is a rise in G and a fall in t, the tax rate).

In Figure 13.14, we show what happened in France and in its biggest trading partner, Germany. The blue bars show the outcomes for France and the red bars show the outcomes for Germany. The figure presents the outcomes for three years: in the first, there was no stimulus, in the second, there was a fiscal stimulus in France and the third year was the year following the stimulus.

If you look at Figure 13.14, you will see that the budget balance in France (measured as $(T - G)/Y$) becomes negative. We can read this as saying that from a balanced budget in 1980, there was a budget deficit of nearly 3% of GDP in 1982, which increased to more than 3% of GDP by 1983.

Meanwhile, in Germany, the budget remained close to balance through the three years. The budget surpluses were 0%, 0% and 0.2%.

The expansionary demand policy in France was an exception in Europe. There was an initial boost to French growth in 1982 (from 1.6% to 2.4%) but it quickly vanished, with growth falling back to 1.2% in 1983. Why?

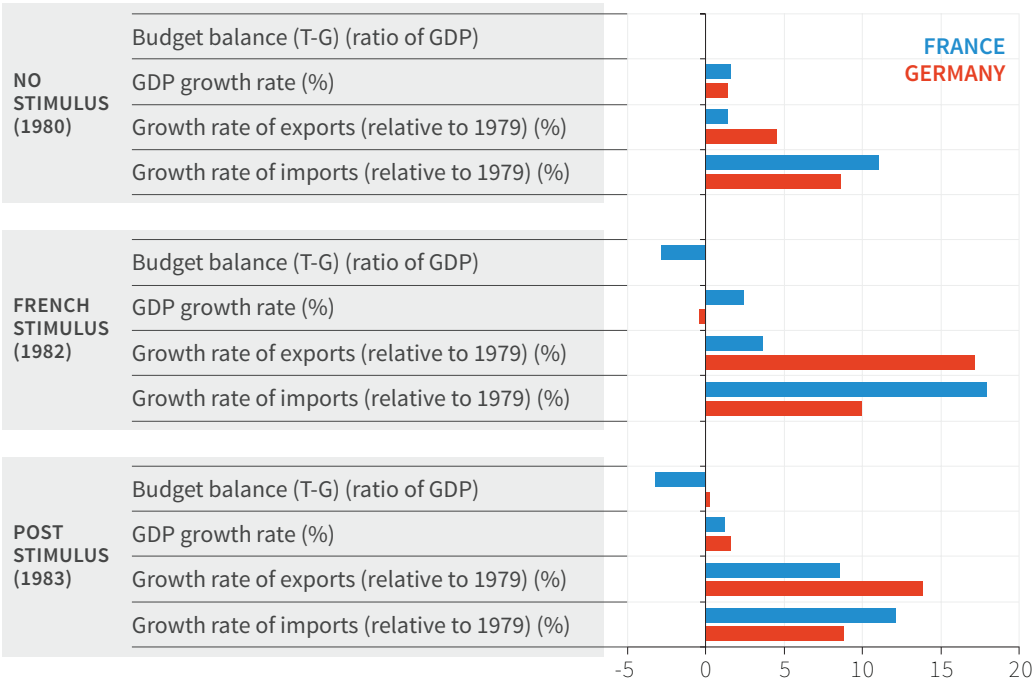


Figure 13.14 *Successes and failures of the French fiscal stimulus (1980-1983).*

Source: *OECD. 2015. 'OECD Statistics.'*

The upturn in the French economy led French households to increase their spending; but much of this was on foreign goods. The French stimulus spilled over to countries producing more competitive products, like Japanese electronic goods and German cars. There was a surge of imports into France: measured relative to the level in 1979, imports were higher by 17.9%, as shown in Figure 13.14. Germany's exports were higher by 17.1% in 1982 and by nearly 14% in 1983. As a result, GDP growth was higher in Germany than in France in 1983. The French stimulus policy mostly benefitted its trading partners with more competitive goods. France slipped behind the pack of European countries, with lower growth and a high government budget deficit (above 3% in 1983).

The failure of Mitterrand's policy was reflected in economic terms by pressure on the French franc (the unit of currency during the period): between 1981 and 1983, the French government had to devalue three times in an effort to make French goods more competitive with those produced abroad. Mauroy stepped down in 1983 and the new prime minister introduced an austerity policy.

The Mitterrand experiment highlights the limits to successful stabilisation of a deep recession by the use of fiscal stimulus. In the case of France, the policy was badly designed and it delayed the adjustment of the French economy to the shocks that had affected it in the 1970s. Note that the problem in France was not only high unemployment: injecting more aggregate demand stimulated spending, but not spending on French output.

The multiplier was very low; the spillover effects to other economies meant that most of the stimulus leaked out of France. Had the major European economies adopted fiscal expansionary policies simultaneously the results would have been different, as the spillover effects of Germany, say, would have stimulated the French economy. This is an example of poor policymaking. It would fit in the bottom right-hand box of Figure 13.12.

DISCUSS 13.10: COORDINATING A STIMULUS

Assume the world is made up of just two countries, or blocs, called North and South. The world is in a slump. The situation can be described using the coordination game used for investment in Unit 12. Here the two strategies are *Stimulus* and *No stimulus*.

Explain in words how the coordination game reflects the problems faced by policymakers in the two countries that arise because of their interdependence.

13.10 AGGREGATE DEMAND AND UNEMPLOYMENT

We now have two models for thinking about total output, employment, and the unemployment rate in the economy:

- *The supply side:* One model, set out in Unit 9, is of the supply side of the economy and focuses on how labour is employed to produce goods and services. This is called the labour market model (or the wage curve and profit curve model).
- *The demand side:* The other is of the demand side of the economy and explains how spending decisions generate demand for goods and services and, as a result, employment and output. This is the multiplier model.

When we put the models together we will be able to explain how the economy fluctuates around the long-run labour market equilibrium over the business cycle.

The labour market model from Unit 9 is shown in Figure 13.15, and the equilibrium in the labour market is where the wage and profit curves intersect. We will see that the economy tends to fluctuate over the business cycle around the unemployment rate shown at point A. In the example in Figure 13.15, the unemployment rate at equilibrium is 5%.

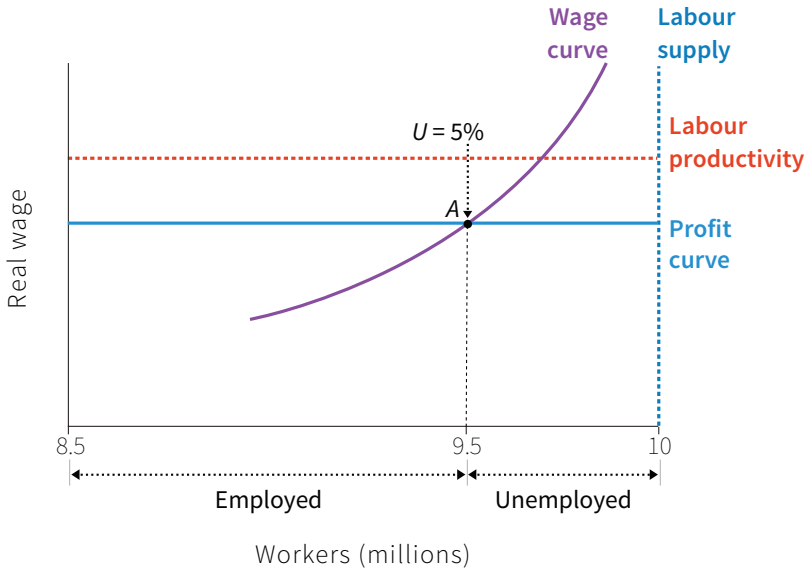


Figure 13.15 *The supply side of the aggregate economy: The labour market.*

Figure 13.16 places the multiplier diagram beneath the labour market diagram. Note that in the labour market diagram, the number of workers is on the horizontal axis and we can measure employment and unemployment along it. In the multiplier diagram, output is on the horizontal axis. The *production function* connects employment and output and, in this model, the production function is very simple.

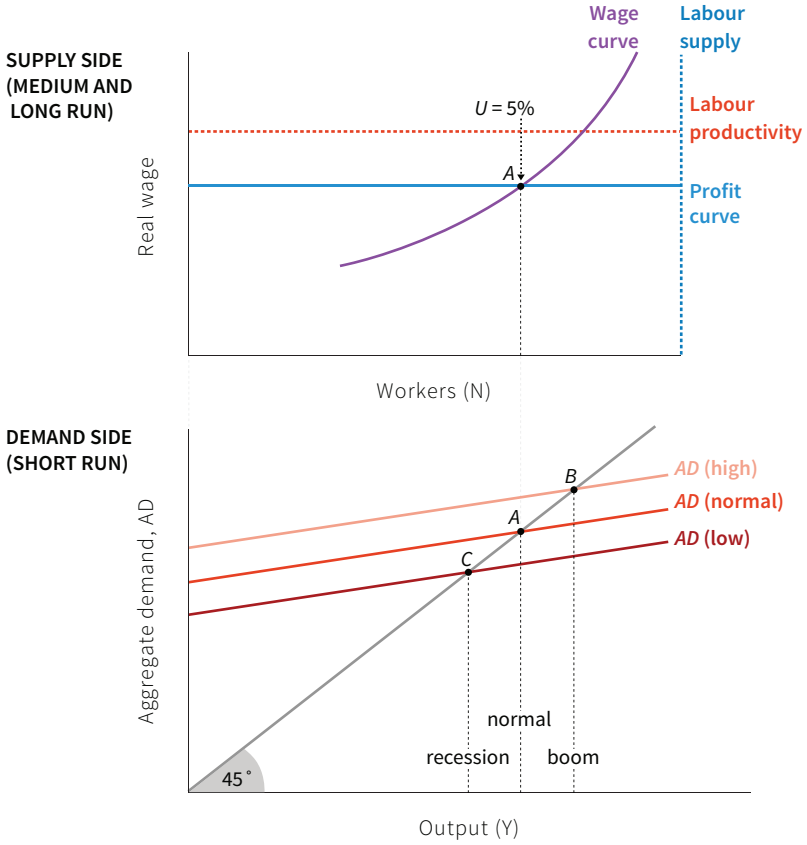


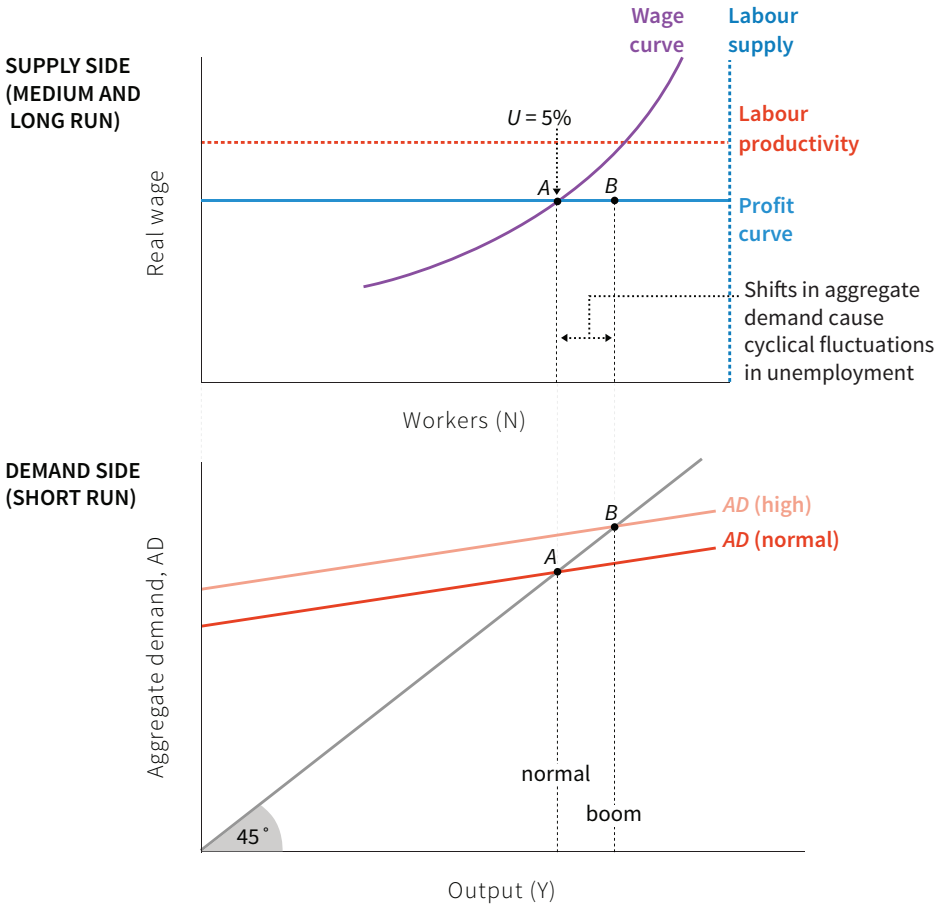
Figure 13.16 *The supply side and demand sides of the aggregate economy.*

Labour productivity is constant and equal to λ ("lambda"), so the production function is:

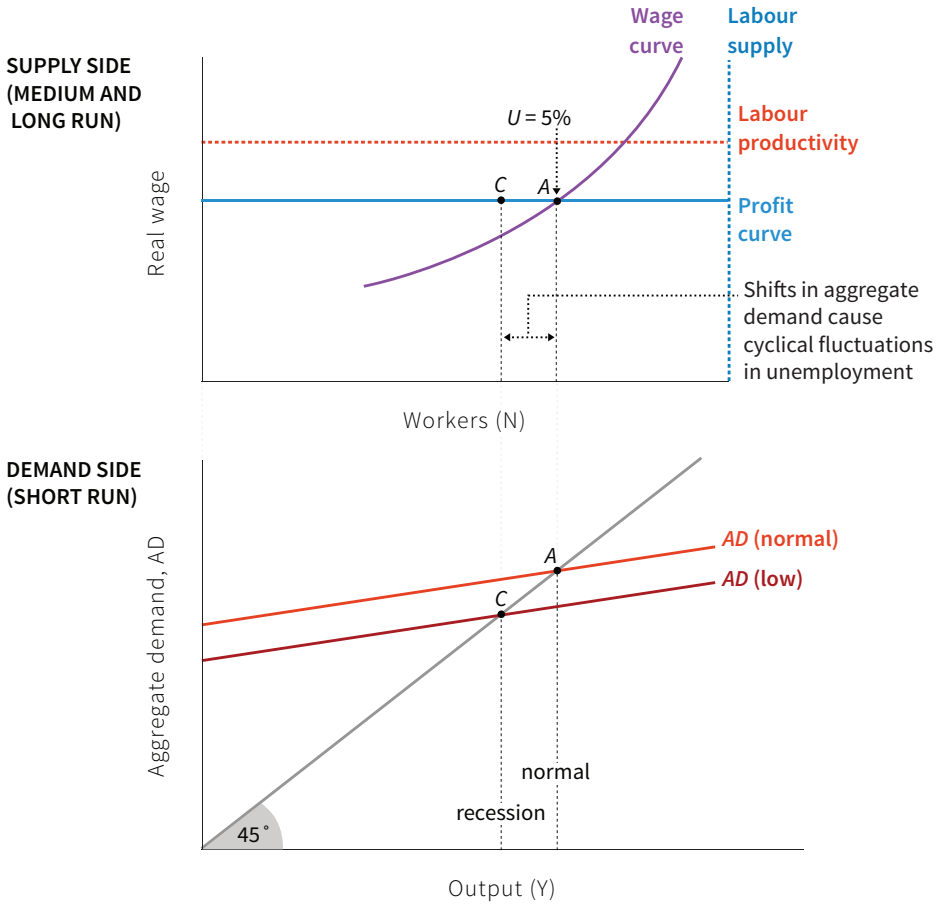
$$Y = \lambda N$$

To allow us to draw the demand-side model underneath the supply-side model, we assume $\lambda = 1$, and so $Y = N$.

Short-term fluctuations in employment are caused by changes in aggregate demand. In Figure 13.17, the economy is initially at labour market equilibrium at point A with unemployment of 5%. The level of output here is called the *normal level* of output. This means that the level of aggregate demand must be as shown by the aggregate demand curve labelled "normal". Any other level of aggregate demand would produce a different level of employment.



Consider a rise in investment that shifts the aggregate demand curve up to *AD(high)*; output and employment rise. The economy is at *B*: with the boom, unemployment falls below 5%.



If the aggregate demand curve shifts down, then through the multiplier process, output and employment fall to *C*. Unemployment rises above 5%.

Figure 13.17 *Business cycle fluctuations around equilibrium unemployment.*

In our study of business cycle fluctuations using the multiplier model, we have made a number of ceteris paribus assumptions. We have assumed that prices, wages, the capital stock, technology and institutions are constant. We use the term short run to refer to these assumptions. The purpose of the model is to predict what happens to output, aggregate demand and employment when there is a demand shock (a shock to investment, consumption or exports), or when policymakers use fiscal policy or monetary policy to shift the aggregate demand curve.

Figure 13.18 summarises the different models we will use to study the aggregate economy.

UNIT	RUN	WHAT IS EXOGENOUS?	WHAT IS ENDOGENOUS?	PROBLEM TO BE ADDRESSED	APPROPRIATE POLICIES	MODEL TO USE
12, 13	Short	Prices, wages, capital stock, technology, institutions	Employment, demand, output	Demand shifts affect unemployment	Demand-side	Multiplier
13, 14	Medium	Capital stock, technology, institutions	Employment, demand, output, prices, wages	Demand and supply shifts affect unemployment, inflation and equilibrium unemployment	Demand-side, supply-side	Labour market; Phillips curve
15	Long	Technology, institutions	Employment, demand, output, prices, wages and capital stock	Shifts in profit conditions and changes in institutions affect equilibrium unemployment and real wages	Supply-side	Labour market model with firms neither entering nor leaving

Figure 13.18 *Models to study the aggregate economy..*

Notice that in figure 13.17 the labour market is not in equilibrium when output is higher or lower than normal. The labour market model is a *medium-run* model where wages and prices can change, unlike in the multiplier model, which is a *short-run* model. So a short-run equilibrium in the multiplier model may not be a medium-run equilibrium in the labour market model.

- *In Unit 14:* We develop the model in Figure 13.17 by asking what happens to wages and prices in a boom and in a recession.
- *In Unit 15:* We use the wage curve and the profit curve to study the *long run*, where not only output, employment, prices and wages can change, but also institutions and technologies. We ask how changes in basic institutions and policies such as the weakening of trade unions, the increase in competition in markets for goods and services, or new labour-saving technologies will affect the aggregate economy.

13.11 CONCLUSION

President Kennedy's crash course in economics did not teach him to worry about the problems that President Mitterrand would encounter, but he did learn from Harris and the others that government programs, especially those, like unemployment insurance, that would sustain high levels of demand even during economic downturns, could moderate economic instability and avoid the very high levels of unemployment that had afflicted the US and other capitalist economies in the Great Depression.

The performance of the US economy was extraordinarily good until a decade after Kennedy's assassination in 1963. The US had limited unemployment, rapid economic growth and little inflation. We can debate whether Harris (or Keynes) should get the credit. The high levels of government spending during the US war in Vietnam, high levels of investment stimulated by the expansion of the economy itself, and rapid increases in labour productivity also contributed to this golden age.

In other countries, too, the expansion of the government sector even without the stimulus of war spending coincided with a dramatic stabilisation of the economy compared to earlier decades. We look more closely at this golden age in Unit 17.

CONCEPTS INTRODUCED IN UNIT 13

Before you move on, review these definitions:

- *Multiplier* mechanism
- *Marginal propensity to consume, marginal propensity to import*
- *Consumption function*
- *Investment function*
- *Goods market equilibrium*
- *Autonomous consumption, Autonomous demand*
- *Target wealth*
- Household *balance sheet*
- *Financial accelerator*
- *Automatic stabiliser*
- *Fiscal stimulus*
- *Paradox of thrift*
- *Government budget balance, deficit, surplus; Primary deficit, Government debt*
- *Positive* and *negative feedback*
- *Supply* and *demand sides* of aggregate economy
- *Business cycle* fluctuations
- *Long run, Medium run, Short run*

But the years since then have taught policymakers and economists alike that it is a challenge to manage unemployment and inflation at the same time. We take up this challenge in the next unit.

Key points in Unit 13

The multiplier

A change in aggregate demand will change output more than one-for-one when households use some portion of any increase in income they receive to purchase goods and services. Because the multiplier is typically greater than one, positive or negative shocks to aggregate demand are amplified.

Government transfers are stabilisers

Transfers such as unemployment benefits typically increase in a recession and tax revenues tend to fall. These automatic stabilisers help to dampen the business cycle.

Fiscal and monetary policy

The government and central bank can use fiscal policy (changes in taxes or spending) or monetary policy to stabilise the economy. Fiscal policy affects aggregate demand directly, whereas monetary policy affects aggregate demand indirectly by altering interest rates.

Fiscal stimulus can stabilise the economy

This can be used when private investment or consumption falls, or when the private sector increases its saving. It needs to be reversed once private spending recovers and the economy is growing again in order to avoid government debt rising unsustainably.

Debt relative to GDP

Wars, recessions and financial crises increase government debt relative to GDP. Growth of the economy, low interest rates and inflation reduce it.

Trade with the rest of the world

This is a source of good and bad shocks to aggregate demand, such as export booms or the collapse of demand in another country's market. Trade with the rest of the world reduces the size of the multiplier.

Calculating the multiplier

We can summarise our findings from the diagram by doing some algebra. To get the multiplier, we can calculate the total increase in production after n+1 rounds of the process described above. Each round of the process matches the circular flow diagram. The first-round increase in demand and production is €1.5bn, triggered by the increase in investment of €1.5bn. The second-round increase in demand and production is $(c_1 \times €1.5bn)$. The third-round increase in demand and production is $c_1 \times (c_1 \times €1.5bn) = (c_1^2 \times €1.5bn)$ and so on.

Following this logic, the total increase in demand and production after n+1 rounds is the total sum:

$$1.5 + c_1(1.5) + c_1^2(1.5) + \ \ldots + c_1^n(1.5) = 1.5(1 + c_1 + c_1^2 \ + \ldots + c_1^n)$$

Taking account of many rounds of indirect effects (large n), because the marginal propensity to consume is lower than one, we can show that the total sum in the brackets reaches a limit of $1/(1 - c_1)$. This is because the term in the brackets is, mathematically, a geometric series. We show this as follows.

If k is the multiplier, we have:

$$k = (1 + c_1 + c_1^2 + \ldots + c_1^n)$$

where we think of n as some large number. Now multiply both sides by $(1 - c_1)$ to get:

$$\begin{aligned} k(1 - c_1) &= (1 + c_1 + c_1^2 \ + \ldots + c_1^n)\,(1 - c_1) \\ &= (1 + c_1 + c_1^2 \ + \ldots + c_1^n) - (c_1 + c_1^2 \ + c_1^3 + \ldots + c_1^{n+1}) \\ &= 1 - c_1^{n+1} \end{aligned}$$

Now divide again by $(1 - c_1)$:

$$k = \frac{(1 - c_1^{n+1})}{(1 - c_1)}$$

As n gets large, assuming $c_1 < 1$, the numerator tends to 1. So, in the limit:

$$k = \frac{1}{(1 - c_1)}$$

In the example, the marginal propensity to consume is, on average, 0.6. This implies that the multiplier is equal to:

$$\frac{1}{(1-c_1)} = \frac{1}{(1-0.6)} = 2.5$$

We can then multiply the multiplier by the initial change in investment of €1.5bn to find the sum of all the successive increases in production triggered by the initial hike in investment and aggregate demand: *2.5 x €1.5bn = €3.75bn*.

The multiplier in an economy with a government and foreign trade

We can again use the fact that there is equilibrium in the goods market when output is equal to aggregate demand (this is where the aggregate demand line crosses the 45-degree line in the multiplier diagram) to find the multiplier. The equation can be rearranged to solve for output and consequently the multiplier:

output = aggregate demand

output = consumption + investment +government spending + net exports

Therefore:

$$Y = c_0 + c_1(1-t)Y + I(r) + G + X - mY$$

$$Y(1 - c_1(1-t) + m) = c_0 + I(r)G + X$$

$$Y = \underbrace{\frac{1}{(1-c_1(1-t)+m)}}_{\text{multiplier}} \underbrace{(c_0 + I(r)G + X)}_{\text{demand that doesn't depend on income}}$$

We can see that the multiplier is smaller when we introduce the government and foreign trade:

$$\frac{1}{(1-c_1(1-t)+m)} < \frac{1}{(1-c_1)}$$

The reason is that the denominator on the left-hand side is larger than on the right:

$$1 - c_1(1-t) + m > 1 - c_1$$

13.12 READ MORE

Bibliography

1. Acconcia, Antonio, Giancarlo Corsetti, and Saverio Simonelli. 2014. 'Mafia and Public Spending: Evidence on the Fiscal Multiplier from a Quasi-Experiment.' *American Economic Review* 104 (7): 2185–2209.
2. Almunia, Miguel, Agustín Bénétrix, Barry Eichengreen, Kevin H. O'Rourke, and Gisela Rua. 2010. 'From Great Depression to Great Credit Crisis: Similarities, Differences and Lessons.' *Economic Policy* 25 (62): 219–65.
3. Auerbach, Alan, and Yuriy Gorodnichenko. 2015. 'How Powerful Are Fiscal Multipliers in Recessions?' *NBER Reporter 2015 Research Summary*. Number 2.
4. Bank of England. 2015. 'Three Centuries of Macroeconomic Data.'
5. Barro, Robert J. 2009. 'Government Spending Is No Free Lunch.' *Wall Street Journal*, January 22.
6. Blanchard, Olivier. 2012. 'Lessons from Latvia.' *iMFdirect - The IMF Blog*. June 11.
7. DeLong, Bradford. 2015. 'Draft for Rethinking Macroeconomics Conference Fiscal Policy Panel.' *Washington Center for Equitable Growth*. April 5.
8. Federal Reserve Bank of St. Louis. 2015. 'FRED.'
9. Gordon, Robert J. 1986. *The American Business Cycle: Continuity and Change*. Chicago, Il: University of Chicago Press.
10. Harford, Tim. 2010. 'Stimulus Spending Might Not Be as Stimulating as We Think.' *Financial Times Undercover Economist Blog*, January 9.
11. Harris, Seymour. 1948. *Saving American Capitalism: A Liberal Economic Program*. New York, NY: A.A. Knopf.
12. International Monetary Fund. 2012. *World Economic Outlook October: Coping with High Debt and Sluggish Growth*.
13. Keynes, John Maynard. (1936) 1997. *The General Theory of Employment, Interest, and Money*. Amherst, NY: Prometheus Books.
14. Keynes, John Maynard. (1926) 2004. *The End of Laissez-Faire*. Amherst, NY: Prometheus Books.
15. Keynes, John Maynard. (1919) 2005. *The Economic Consequences of Peace*. New York, NY: Cosimo Classics.
16. Krugman, Paul. 2009. 'War and Non-Remembrance.' *The New York Times*, January 22.
17. Krugman, Paul. 2012. 'A Tragic Vindication.' *The New York Times*, October 9.
18. Krugman, Paul. 2014. 'The Fall of France.' *The New York Times*, November 6.

19. Leduc, Sylvain, and Daniel Wilson. 2015. 'Are State Governments Roadblocks to Federal Stimulus? Evidence on the Flypaper Effect of Highway Grants in the 2009 Recovery Act.' *Federal Reserve Bank of San Francisco Working Paper* 2013-16 (September).
20. Muellbauer, John. 2014. 'Combatting Eurozone Deflation: QE for the People.' *VoxEU.org*. December 23.
21. OECD. 2015. 'OECD Statistics.'
22. Portes, Jonathan. 2012. 'What Explains Poor Growth in the UK? The IMF Thinks It's Fiscal Policy.' *National Institute of Economic and Social Research Blog*. October 9.
23. Romer, Christina D. 1993. 'The Nation in Depression.' *Journal of Economic Perspectives* 7 (2): 19–39.
24. Samuelson, Paul A. 1998. *Economics: The Original 1948 Edition*. New York, NY: McGraw-Hill.
25. Shiller, Robert J. 2015. 'Economic Stimulus, without More Debt.' *The New York Times*, January 9.
26. Skidelsky, Robert. 2010. *Keynes: The Return of the Master*. London: Penguin.
27. Smith, Noah. 2013. 'Why the Multiplier Doesn't Matter.' *Noahpinion*. January 7.
28. Temin, Peter, and David Vines. 2014. *Keynes: Useful Economics for the World Economy*. Boston, MA: MIT Press.
29. *The Economist*. 2009. 'A Load to Bear.' November 26.
30. *The Economist*. 2011. 'The Maths behind the Madness.' November 9.
31. The Maddison Project. 2013. '2013 Version.'
32. US Bureau of Economic Analysis. 2015. 'GDP & Personal Income.'
33. Wallis, John Joseph. 2000. 'American Government Finance in the Long Run: 1790 to 1990.' *Journal of Economic Perspectives* 14 (1): 61–82.
34. Wren-Lewis, Simon. 2012. 'Multiplier Theory: One Is the Magic Number.' *Mainly Macro*. August 24.

Beta January 2016 version

14

INFLATION AND MONETARY POLICY

HOW THE RATE OF UNEMPLOYMENT AND THE LEVEL OF OUTPUT IN THE ECONOMY AFFECT INFLATION, THE CHALLENGES THIS POSES TO POLICYMAKERS, AND HOW THIS KNOWLEDGE CAN SUPPORT EFFECTIVE POLICIES TO STABILISE EMPLOYMENT AND INCOMES

- When unemployment is low, inflation tends to rise and when unemployment is high, inflation falls
- Policymakers and voters prefer low unemployment and low inflation (but not a falling price level). They typically cannot have both and instead they face a trade-off
- There is an inflation-stabilising rate of unemployment, and a wage-price inflation spiral develops if unemployment is kept lower than this
- Monetary policy affects aggregate demand and inflation through a variety of channels
- An oil price increase, or some other adverse shock, can lead to higher unemployment and higher inflation
- Many governments have given responsibility for monetary policy—often described as inflation targeting—to central banks

Institute for New Economic Thinking
UCL SciencesPo Azim Premji University
FRIENDS PROVIDENT Foundation

Before his successful 1992 US presidential campaign, Bill Clinton's electoral strategists had decided that two of their campaign issues should be health policy and "change". But it was the third focus of his campaign—the recession of 1991—that resonated with the public. The reason was the phrase the campaign workers used: "It's the economy, stupid!"

The 1991 recession meant that many Americans lost their jobs, and the Clinton campaign slogan brought this issue to the attention of the voters. When the ballots were counted in November 1992, Clinton received almost 6 million more votes than George H.W. Bush, the sitting president.

In a democracy election outcomes are always affected by the state of the economy, and how the public judges the economic competence of the government and the opposition. Two important features of this economic performance are unemployment and inflation. In Unit 12, we saw that unemployment undermines our wellbeing, but inflation worries us too. Figure 14.1 shows that, in US presidential elections, the margin of victory of the ruling party is higher when inflation is lower.

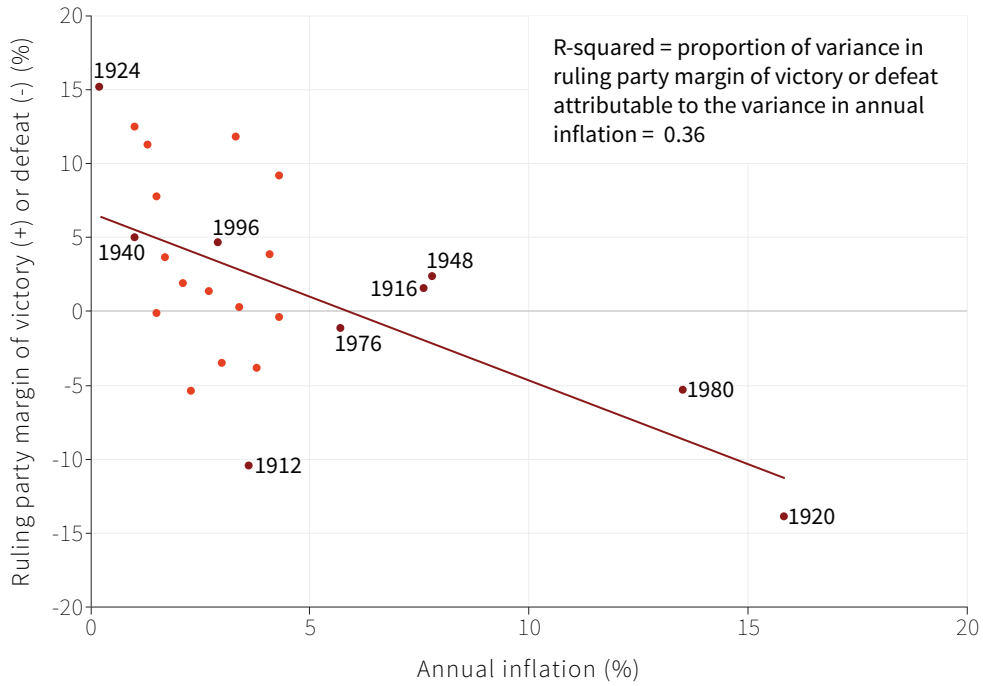


Figure 14.1 *Inflation and presidential elections in the United States (1912-2012).*

Source: *Inflation before 1950: Bordo, Michael, Barry Eichengreen, Daniela Klingebiel, and Maria Soledad Martinez-Peria. 2001. 'Is the Crisis Problem Growing More Severe?' Economic Policy 16 (32): 52–82. CPI after 1950: Federal Reserve Bank of St. Louis. 2015. 'FRED.' Electoral results: US National Archives. 2012. '1789-2012 Presidential Elections.' US Electoral College.*

So, if you were a politician worrying about your citizens' concerns as well as your own career, you should minimise both unemployment and inflation. Is this possible? We get an insight by looking at how a German minister of finance, trained as an economist, handled his dual role as a politician (at an election rally in the evening) and as an economist (in his office the next day).

Helmut Schmidt was called the "superminister" in the West German government of Chancellor Willy Brandt because he was both minister of economics and minister of finance. At an election rally in 1972, he claimed that it would be possible to keep unemployment below 5% without inflation rising above 5%. He was promising that his party would choose a combination of inflation and unemployment that would favour lower unemployment.

The following day Professor Otto Schlecht, head of the economics policy department at the Federal Ministry of Economics, said to Schmidt: "Herr Minister, what you said yesterday, which is in the newspapers this morning, is false."

Schmidt replied:

"I agree that what I said was technically wrong. But you cannot advise me about what I decide is politically expedient to say to an election rally in front of 10,000 Ruhr miners in the Westfalenhalle in Dortmund."

Helmut Schmidt's 5% commitment, and his explanation afterward, show two things about the relationship between economics and politics. The first is that politicians are elected to office, and so respond to the views of voters. The second is that politicians as policymakers face constraints on their choice of policies. They can't just promise the economic outcomes voters care about—in Schmidt's case, low unemployment and low inflation. The economist in Schmidt was well aware of the constraints but, at the rally, he was speaking as a politician.

While the policymaker wants to deliver low unemployment and low inflation, the economy operates in such a way that when unemployment goes down, inflation tends to go up. And when inflation falls, unemployment goes up. This is a problem we have seen before: policymakers must deliver what is feasible, and this involves trading one objective off against the other. Another way to say this: more inflation is the *opportunity cost* of lower unemployment, and more unemployment is the opportunity cost of less inflation. Moreover, the economy is subject to shocks that can make both inflation and unemployment worse, limiting the set of feasible outcomes. And experience from the late 1960s showed that inflation would carry on rising if unemployment were too low. This was the setting for Helmut Schmidt's reflections on his election promise.

Following the experience of rising inflation across the world during the late 1980s, there was a rethinking of how macroeconomic policy should be designed and, in the 1990s, the widespread adoption of the policy known as inflation targeting by central banks. Governments delegated the management of fluctuations in the economy to

the central bank, with fiscal policy playing a lesser role, and recognised that policies to improve the supply side of their economies—such as increasing competition and better functioning labour markets—were necessary if they wanted to achieve a lower rate of unemployment compatible with low and stable inflation.

As we saw in Unit 9, prices are messages: they send signals about scarce resources. We looked at how shifts in demand or supply for a good resulted in a change in its price relative to other goods and services, and how this signalled a change in the relative scarcity of the good or service. In this unit we look not at relative prices but at inflation or deflation: a rise or fall in prices in general. We begin by asking how inflation got a bad name.

14.1 WHAT'S WRONG WITH INFLATION?

Before we turn to the question, we need to clarify a few terms.

What is the difference between *inflation* and rising inflation? What is *deflation*?

A useful way to think about these differences is to compare what happens to the price level in the economy with a car's initial location and the distance covered when it travels at different speeds.

Zero inflation

A constant price level from year to year means that inflation is zero. This can be compared with a stationary car: the car's location is constant and the distance travelled per hour is zero.

Inflation

Now, consider a rate of inflation, such as 2% per year. This means that the price level goes up by 2% each year. In the case of the car, a car travelling at 2km per hour means that the distance from the initial location increases by 2km each hour. After two hours, the car is 4km away from its initial location; after another hour, it is 6km distant, and so on.

Deflation

Deflation is when the price level falls. If the price level falls by 2% per year, prices a year later are 2% lower than initially. In the case of the car, this is equivalent to the car travelling backward at 2km per hour. After an hour, the car is 2km behind its initial location, and so on.

Rising inflation

DESCRIBING A CHANGE IN PRICE LEVEL

- *Inflation*: The price level is rising
- *Deflation*: The price level is falling
- *Disinflation*: The *inflation rate* is falling

Suppose now that the rate of inflation increases from 2% to 4% to 6% in successive years: the economy experiences rising inflation. The equivalent for the car is that speed picks up from 2km per hour in the first hour, to 4km per hour in the second hour, and so on. After two hours, the car is 6km away from its initial location. We talk of the car accelerating: the distance travelled from the starting point is increasing at an increasing rate. Similarly, in the economy, if the rate of inflation is increasing, the price level is increasing at an increasing rate.

Falling inflation

This is called disinflation and is equivalent to a car reducing its speed, for example from 6km per hour, to 4km per hour, to 2km per hour. Once the speed reaches zero, the car's location does not change. The equivalent in the economy is that when inflation falls to zero, the price level does not change.

We have seen why voters dislike unemployment. But why do voters dislike inflation? Think of pensioners or others in the economy whose income is fixed in nominal terms. This means it is fixed in terms of yuan or dollars or euros. If *prices rise* during the year, such households can buy fewer goods and services at the end of the year than they could at the beginning. They are worse off and will tend to vote against a party they believe will permit higher inflation.

Whether one loses or benefits from inflation also depends on which side of the credit market one is on. Julia the borrower and Marco the lender (in Unit 11) have a conflict about the interest rate at which Julia borrows. They also have differing interests about inflation, because if prices rise before Julia repays her loan, Marco will find that he can buy less with the repayment than would have been the case if there was zero inflation.

More generally, using the same logic as we used when discussing the government's debt in the previous unit, inflation means that:

- *People with nominal debt will benefit*: Those with mortgages on fixed *nominal interest rate* loans, for example, will benefit from inflation, because the debt stays the same in nominal terms, and so becomes smaller in real terms.
- *People with nominal assets will lose:* Those people with deposits in banks, for example. The same is true of banks or others who have loaned money at interest rates fixed in nominal terms, because when the sum is repaid it will be worth less in terms of the goods or services it can buy. Very high inflation will wipe out the value of nominal assets, as happened in Zimbabwe.

To take account of inflation when analysing borrowing and lending we use what is termed the *real interest rate*, which is defined as follows:

real interest rate (% per annum) = nominal interest rate (% per annum) – the inflation rate (% per annum)

The real interest rate measures the buying power of the repayment of a loan at the prices that exist when the loan is repaid. To see what this means, let's suppose Julia were to borrow $50 from Marco with a repayment of $55 next year. The nominal interest rate is 10%. But if next year's prices were 6% higher than this year's (inflation of 6%) then what Marco could buy with the repayment is not 10% more than he could have bought with the sum he loaned to Julia, but instead only 4%. The real interest rate is 4%.

In addition to redistributing income from *creditors* (those with assets) and those on nominally fixed incomes (like pensioners) to *debtors*, in some cases inflation can also make the economy work less well. While there is no evidence that moderate inflation is bad for the economy, when inflation is high it is often also volatile and therefore hard to predict. Large price changes create uncertainty, and make it more difficult for individuals and firms to make decisions based on prices.

In an environment of high and volatile inflation, it is hard to separate the signal about the scarcity of resources (sent by *relative prices*) from the noise of erratically rising prices caused by an inflationary environment in which the *price level* changes. Firms might find it harder to know in which sector to invest, or which crop to plant (quinoa or barley, for example); individuals would find it harder to decide whether quinoa has become more expensive relative to other sources of protein. Moreover, in an inflationary environment, firms have to update their prices more frequently than they would prefer. This incurs costs in management time, referred to as *menu costs*.

Would households and firms be better off with falling prices? No. A sustained fall in the price level is undesirable for many of the same reasons that inflation is undesirable, and could have even more dramatic economic consequences. When prices are falling, households will postpone consumption (particularly of expensive items such as fridges, screens and cars) because they expect goods will be cheaper in the future. Similarly, deflation increases the debt burden of borrowers, for the same reason that inflation reduces it.

As we have seen in Unit 13, a rise in the debt burden depresses consumption because some affected households save to restore their target wealth and others find themselves credit-constrained. The fall in consumption will induce a drop in aggregate demand and economic activity. Weaker aggregate spending will tend to depress prices further and can trigger a vicious circle of falling prices and economic stagnation.

This happened in Japan. The Japanese economy was one of the great success stories of the period after the second world war: the upward slope of its hockey stick was remarkably steep, as you saw in Unit 1. Living standards, as measured by GDP per capita, went from less than one-fifth of the level in the US in 1950 to more than 70% by 1980. But in the past 25 years low growth and rising unemployment have become entrenched. For the first time in an advanced economy in the post-war period, there has been persistent deflation. Deflation was observed in 11 years out of 20 between 1995 and 2014.

Many economists think that a little bit of inflation (as long as it remains constant) is a good thing. We will see in the next unit why this is the case: the process of innovation and change that characterises a dynamic economy means that, in any given year, there will be losers as well as winners. With rising prices, a fall in real income among the losers may be masked by the fact that nominal incomes are rising or at least not falling. For example, many people will not even notice a slight fall in their real wage due to modest inflation, but nobody would fail to notice a reduction in his or her nominal wage. With some low inflation, the adjustment of workers and resources between different firms and industries in response to changes in relative prices can take place without losers experiencing falling nominal wages. Inflation greases the wheels of the labour market.

14.2 INFLATION RESULTS FROM CONFLICTING AND INCONSISTENT CLAIMS ON OUTPUT

Inflation arises from conflicts among economic actors when they are sufficiently powerful that their claims on goods and services are inconsistent.

To see how this works, think of an economy composed of many firms (each of which is owned by a single individual) and their employees, who are also the consumers of the various goods produced by the firms. To keep track of what is happening in the firms, we assume that prices are set by the marketing department and wages by the human resources (HR) department.

Initially the marketing department in each firm is setting prices based on the markup that maximises its profits given the degree of competition in the markets in which it sells (as we saw in Unit 7 and Unit 9). And the HR department is also setting the real wage for its workers (which is the nominal wage in the firm, divided by the price level in the economy) as the lowest wage consistent with workers actually working, given the level of unemployment in the economy (as we saw in Unit 6 and Unit 9).

If, once all firms have set their wages and prices, the wage rate and the price level are consistent with the firms maximising their profits, then there will be no reason for either prices or wages to be changed. At this unemployment rate, the price level is constant (inflation is zero). This is the level of unemployment where the wage and profit curves intersect, that is, labour market equilibrium.

Suppose now that the government adopts *protectionist policies* making it difficult for foreign firms to enter its markets. Then the markets facing the firm become less competitive, so that the firm can charge a higher markup on its costs. If this is the case across the economy, the resulting increase in the price level will lower the real wage of the workers. But while the owner of an individual firm is happy with the higher price that the marketing department can now charge, the workers are unhappy with the fall in the real wage: the result is that workers now lack the motivation to work. So the HR department of the firm will raise its nominal wage; and all other firms will do the same. So both prices and wages have risen. The economy experiences inflation.

Will it end there? No. The nominal wage increase has raised the cost of production to firms and they will use this as the basis of their markup pricing, leading to a further increase in prices and a fall in the real wage, which the HR department will correct by again raising the nominal wage. The process of rising wages and prices will continue as long as:

- *Firms are powerful enough* to charge the higher markup
- *Workers at the given unemployment rate are powerful enough* to require the initial real wage in order to motivate them to work

In the example given, inflation rose while unemployment did not change, following a change in the competitive conditions facing firms that allowed them to raise their markup, increasing the owners' profits. But there are other ways that the process could have begun from the same starting point. Suppose the degree of competition in product markets remains the same, but the level of employment rises. At the new lower level of unemployment the firms would want to pay workers a higher real wage to keep them working. This induces the marketing departments of firms to raise their prices, so as to maintain the markup that competitive conditions allowed. And the inflationary process would begin.

To summarise, inflation may result from:

- *An increase in the bargaining power of firms over their consumers:* This is caused by a reduction in competition, which allows firms to charge a higher markup. It is a downward shift of the profit curve.
- *An increase in the bargaining power of workers over firms:* This allows them to get a higher wage in return for working hard.

There are two ways that the increase in the bargaining power of workers could take place:

- *A shift upward of the wage curve:* At every level of employment the wage they would receive is higher.
- *An increase in the level of employment:* In this case, the wage curve is unchanged.

We studied reasons for the shift in the wage curve, such as improved generosity of unemployment benefits or stronger trade unions, in Unit 6. The movement along the wage curve, rather than a shift in the curve, is what we will analyse next.

Figure 14.2 summarises three causes of inflation. The third one—higher employment may result in inflation—came to light when economist William (Bill) Phillips published a scatter plot of annual *wage inflation* and unemployment in the British economy. This is shown in Figure 14.3.

GREAT ECONOMISTS

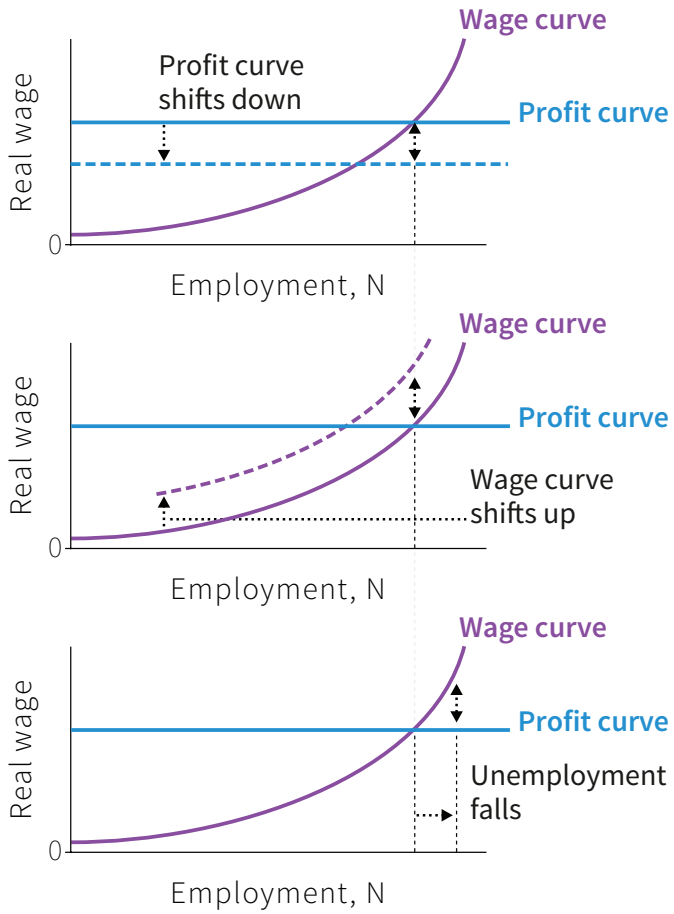
BILL PHILLIPS

William Phillips (1914-1975) was an unusually colourful character for a world-renowned economist. Raised in New Zealand, Phillips spent time as a crocodile hunter, a movie director and a prisoner of war in Indonesia during the second world war, before finally becoming a professor at the London School of Economics.

Phillips had engineering know-how and, while studying sociology in London in 1949, he built a hydraulic machine to model the British economy. It was in the spirit of the hydraulic economy model produced by Irving Fisher half a century earlier, but much more elaborate.

The Monetary National Income Analogue Computer (MONIAC) used transparent pipes and coloured water to bring economists' equations to life. Phillips's machine had tanks for each of the components of domestic GDP, such as investment, consumption and government expenditures; imports and exports were shown by water being added or drained from the model. The machine could be used to model the effect on the economy of shocks to different variables, such as tax rates and government spending, which would set in motion flows between the tanks. Working versions of the machine can still be found in the London Science Museum and universities around the world.

In a 1958 paper, Phillips made another major contribution to the study of economics. By drawing a scatterplot of the data for the rate of unemployment and the rate of wage inflation for the British economy for the years between 1861 and 1913, he revealed an empirical relation between the two variables. He found that lower rates of unemployment were associated with higher rates of inflation and high unemployment with low inflation. The relationship has since been referred to as the *Phillips curve.*



Owner's power rises relative to consumers

For example, lower competition. Medium to long run.

Employees' power rises relative to owners

For example, stronger unions. Medium to long run.

Employees' power rises relative to owners

For example, a business cycle upswing. Short to medium run.

Figure 14.2 *Three causes of inflation.*

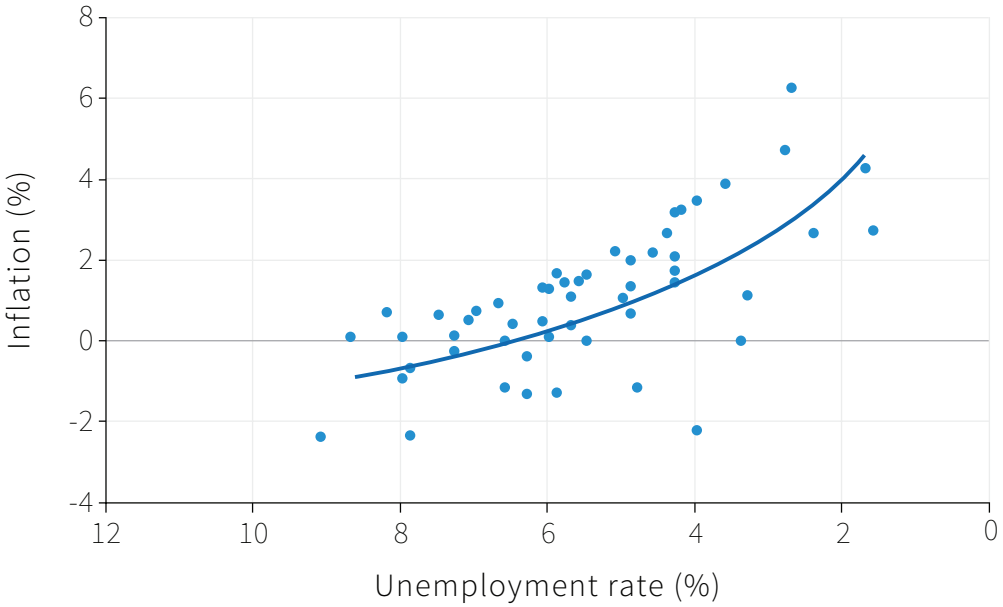


Figure 14.3 *Phillips's original curve: wage inflation and unemployment (1861-1913).*

Source: *Bank of England. 2015. 'Three Centuries of Macroeconomic Data.'*

14.3 INFLATION, THE BUSINESS CYCLE AND THE PHILLIPS CURVE

When central banks report their interest rate decision to the public, they normally justify a rise in the interest rate by saying that *forecast* inflation is up. They are raising the interest rate to dampen aggregate demand, raise unemployment and, as a result, bring inflation back toward target.

Conversely, if they are announcing a lower interest rate, they explain that this is because there is otherwise a danger of inflation falling too low (possibly into deflation). Just as a reduction in aggregate demand and employment will bring inflation down, a rise in aggregate demand and employment will increase inflation.

To model inflation, we assume that the HR departments of firms set nominal wages (for example, in dollars, pounds or euros) once a year, and that the marketing departments set prices immediately after wages. The real wage that employees care about is their nominal wage relative to the economy-wide level of prices, and is:

$$w = \frac{W}{P}$$

It is the *real wage* on the vertical axis in the labour market diagram.

To see how inflation comes about in a business cycle upswing, we begin with constant prices in the economy and consider a rise in aggregate demand, which reduces unemployment.

- *When unemployment is low, the HR department needs to set higher wages:* The cost of job loss is low and workers expect higher real wages to work effectively.
- *Higher wages mean higher costs for firms:* The marketing department will raise prices to cover the higher costs. As long as competitive conditions have not changed, the firm's markup will be unchanged.
- *The price level will have gone up:* Once all firms in the economy have set higher prices, the economy has experienced wage and price inflation. And real wages have not have increased: the percentage increase in W equals the percentage increase in P, so W/P is unchanged.

What happens next? We assume that aggregate demand remains high enough to keep unemployment at its low level. At the next annual round of wage setting, the HR department is in the same position as the previous year: with continuing low unemployment, workers are disappointed with their real wage. It must raise nominal wages. When costs go up, the marketing department raises prices once more. This is called the *wage-price spiral*. It explains why, at low unemployment, the price level rises—not just in the year that unemployment fell—but year after year.

If there is a recession instead of a boom, the wage price spiral operates in reverse, and the price level falls year after year.

We now ask why prices would have been constant year after year before the boom in aggregate demand reduced unemployment. We will see that, when the labour market is in equilibrium (the normal phase of the business cycle) there is no pressure for wages and prices to change. From Unit 9 (and Unit 13), we know that labour market equilibrium is where the wage curve and the profit curve intersect. But why is this unemployment rate so special for the rate of inflation?

In Figure 14.4a, we can see that it is only when the labour market is in equilibrium at point A that the real wage on the wage curve coincides with the real wage resulting from firms' price-setting decisions shown by the profit curve. At A, the claims of owners for profits and of workers for wages add up exactly to the size of the pie (the sum of the double-headed arrows showing the profits per worker and real wages is equal to output per worker, which is shown by the red dashed line). This means that the HR department will have no reason to raise wages, and with no increase in costs, the marketing department will keep prices unchanged. The real wage will remain constant and no one will be disappointed.

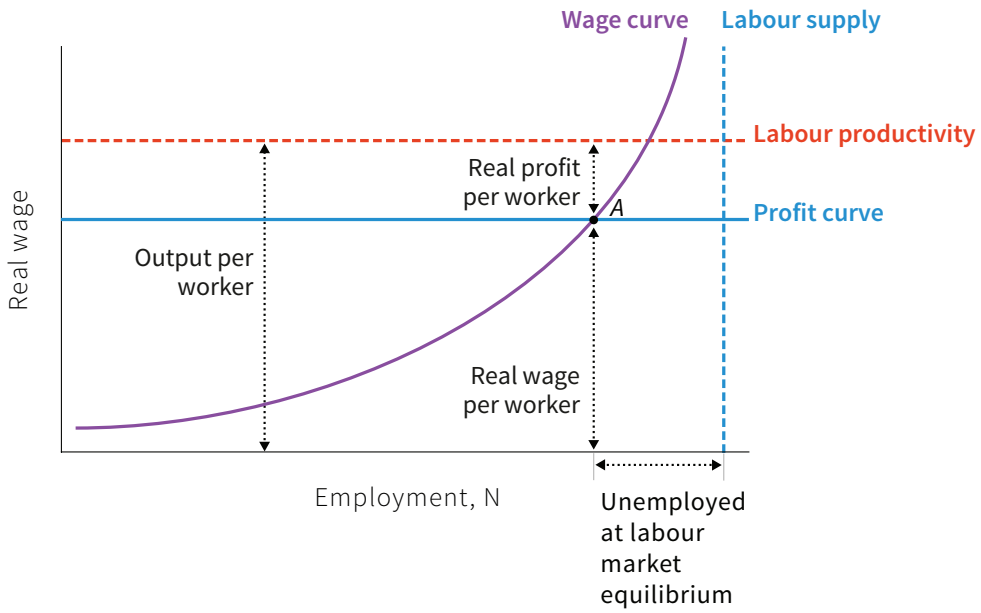


Figure 14.4a *Inflation and conflict over the pie: Stable price level at labour market equilibrium.*

In an economy with stable prices, at the unemployment rate at labour market equilibrium (at *A*), wages and prices will be stable and inflation will be zero.

We now use the labour market diagram to show what happens in a boom, when unemployment is lower than at *A*. Click through Figure 14.4b to check that the real wage required to get workers to work hard increases at lower unemployment (we move up the wage curve to the right). This is why the HR department has to raise nominal wages. And given that the marketing department continues to set prices to deliver unchanged real profits per worker by increasing prices in line with the higher wages, the sum of wage and profit claims exceeds labour productivity at point *B* (the sum of the double-headed arrows is greater than labour productivity shown by the red dashed line). This is an upward wage-price spiral due to inconsistent claims.

In parallel fashion, if unemployment is higher than at *A*, as shown by point *C*, employees are in a weaker bargaining position and by setting lower nominal wages, the HR department still gets adequate effort (we move down the wage curve to the left). Now the sum of wage and (unchanged) profit claims add up to less than labour productivity: there will be downward pressure on wages and prices, a downward wage-price spiral.

If we sketch the relationship between inflation and unemployment from these three phases of the business cycle, we get something similar to the one Phillips discovered in the data: when unemployment is lower, inflation is higher and vice versa.

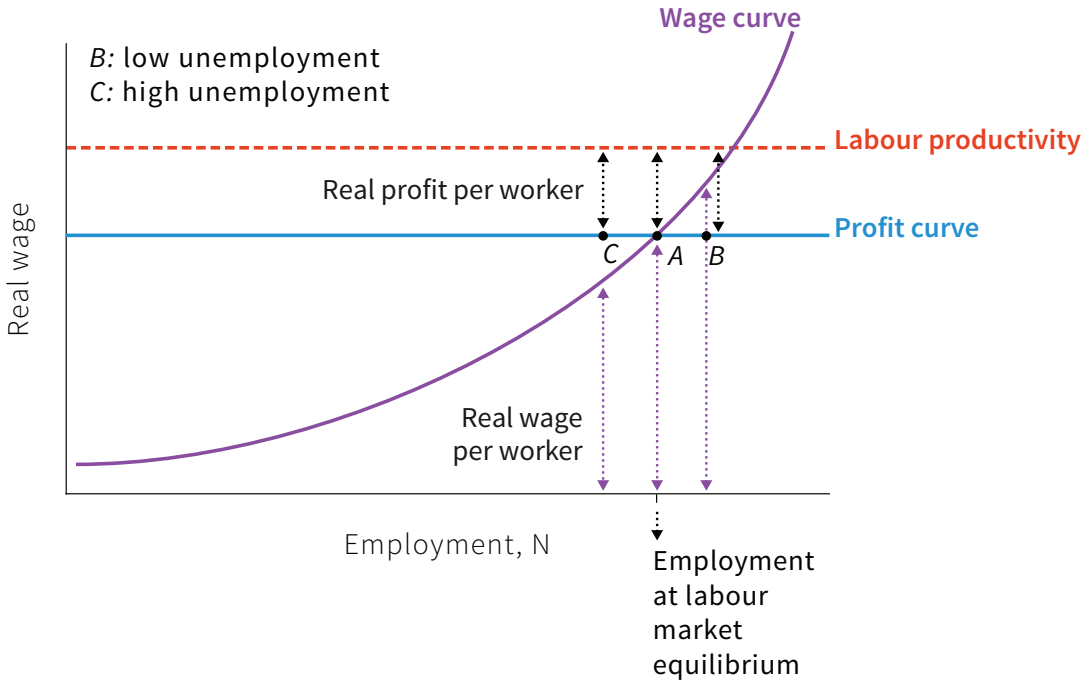


Figure 14.4b *Inflation and conflict over the pie at low and high unemployment.*

The big message from the model of inflation and conflict over the pie is that if employment is so high (or so low) that there is a *bargaining gap*: the wage given by the wage curve and that given by the profit curve are not equal, and so the price level will be either rising or falling.

- *If unemployment is lower than at the equilibrium:* There is a positive bargaining gap and there is inflation.
- *If unemployment is higher than at the equilibrium:* There is a negative bargaining gap and there is deflation.
- *If there is labour market equilibrium:* This is the only situation in which the price level is constant.

For example, if the wage on the profit curve is 100 and on the wage curve it is 101, the bargaining gap is 1%. At labour market equilibrium, the bargaining gap is zero.

BARGAINING GAP

The difference between the real wage that firms wish to offer in order to provide workers with incentives to work, and the real wage that allows firms the markup on costs required to provide owners with the motivation to continue in business.

- When the bargaining gap is positive, the real wage on the wage curve is above that the profit curve, and the claims of employers and owners to output per worker are inconsistent.
- The percentage bargaining gap is equal to the wage on the wage curve, minus the wage on the profit curve, divided by the wage on the profit curve.

The bargaining gap and the Phillips curve

We can summarise the causal chain from the bargaining gap to inflation like this:

Bargaining gap (%) → Increase in wages (%) → Increase in unit costs (%) (assuming no other costs) → Increase in prices (%) ≡ inflation (%)

Remember, the triple bar indicates that inflation is defined as the percentage increase in prices. So, to work out the inflation rate, we use the following:

$$\begin{aligned} \textit{inflation (\%)} &\equiv \textit{increase in prices (\%)} \\ &= \textit{increase in costs per unit of output(\%)} \\ &= \textit{increase in wages (\%) (if wages are the only costs)} \\ &= \textit{bargaining gap (\%)} \end{aligned}$$

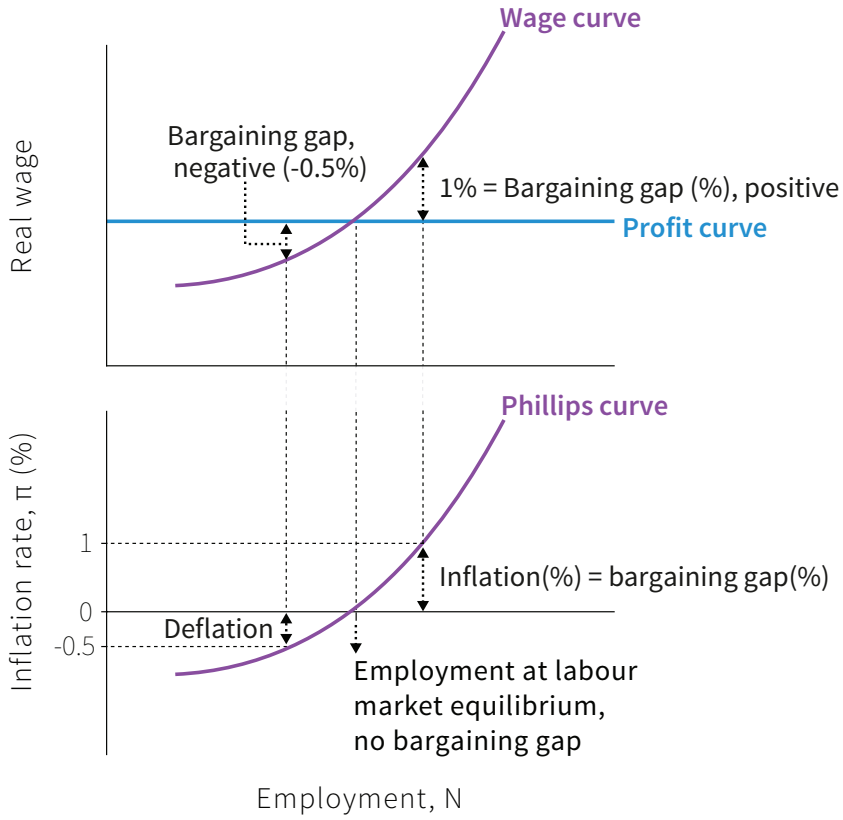


Figure 14.4c *Bargaining gaps, inflation and the Phillips curve.*

In Figure 14.4c, we draw a new diagram beneath the wage curve and profit curve. This is the Phillips curve diagram, with inflation on the vertical axis and employment on the horizontal axis. If we begin with employment at the labour market equilibrium, and inflation of zero, we note that the economy can remain here: there is no pressure for the price level to rise or fall. This gives a point on the Phillips curve. Now consider a higher level of employment due to stronger aggregate demand. A positive bargaining gap opens up and wages and prices will rise. Firms increase wages in

response to the fall in unemployment. The price level rises as firms put up their prices in response to the rise in their labour costs. If the bargaining gap is 1%, prices and wages will rise by 1%. This gives a second point on the Phillips curve.

As long as employment remains above the labour market equilibrium, employees will be disappointed at the end of the year. Their real wage will not have risen by 1% as they had anticipated. The result: wages and prices will rise by 1% the following year as well: firms will put up wages by 1% to take the real wage up to the wage curve, and they will put up prices by 1% in response to that cost increase. We will observe lower unemployment and higher inflation as in Phillips' original empirical scatter plot.

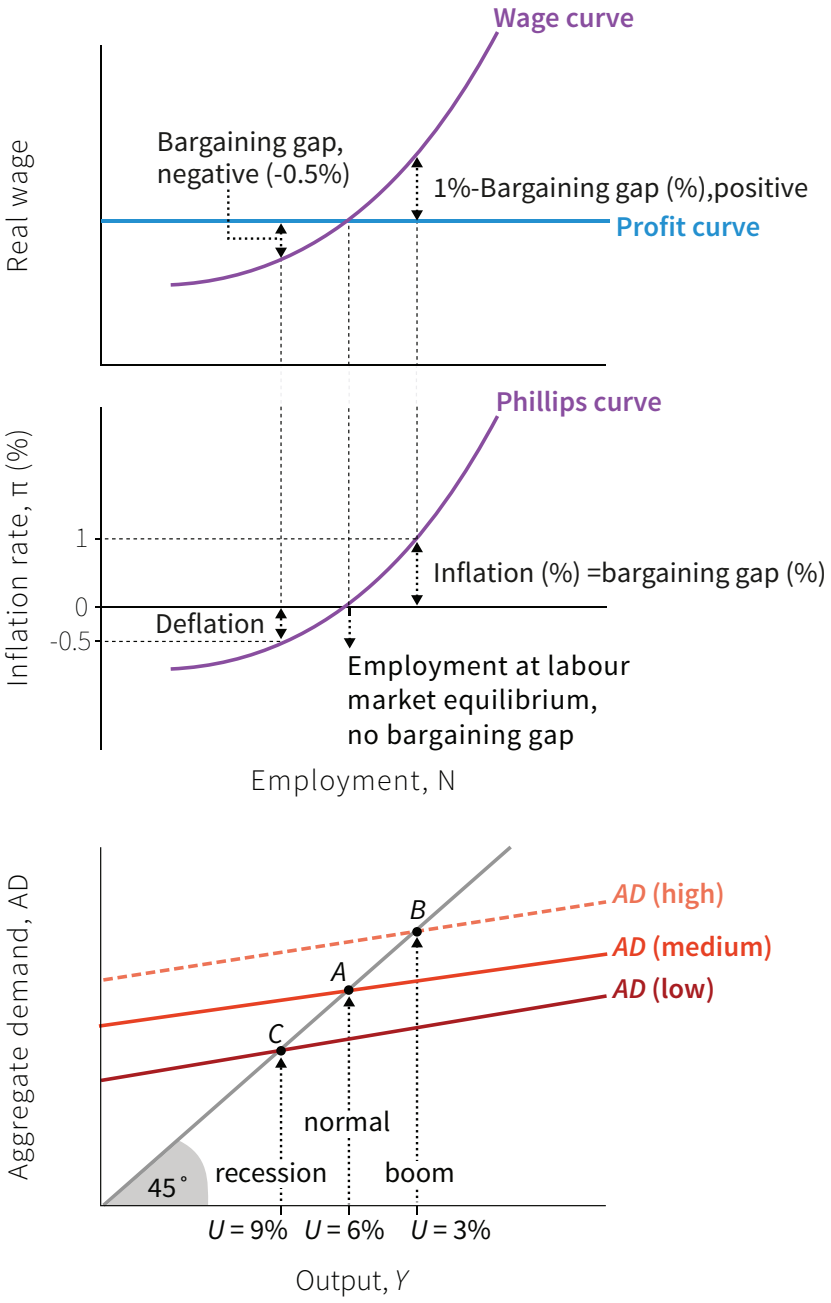


Figure 14.4d *The short- and medium-run models: Aggregate demand, employment and inflation.*

To complete the picture, we include the multiplier model beneath the labour market and Phillips diagrams to bring the short- and medium-run models together. This highlights that:

- *At a higher level of aggregate demand (a boom) inflation is positive:* unemployment is lower, which means there is a positive bargaining gap, and so wages and prices are rising continuously.
- *At a lower level of aggregate demand (a recession), there is deflation:* Unemployment is higher, which means there is a negative bargaining gap.

DISCUSS 14.1: THE BARGAINING GAP IN A RECESSION

Suppose the economy is initially at labour market equilibrium with stable prices (inflation is zero). At the beginning of year 1, investment declines and the economy moves into recession with high unemployment.

1. Explain why a negative bargaining gap arises and assume it is 1%.
2. Draw a diagram with years on the horizontal axis and the price level on the vertical axis. Starting from a price index of 100, sketch the path of the price level for the 5 years that follow, assuming the bargaining gap remains at -1%.
3. Who are the winners and losers in this economy?

DISCUSS 14.2: POSITIVE AND NEGATIVE SHOCKS

Draw a labour market diagram with the economy at labour market equilibrium with stable prices. Now consider:

- A positive shock to aggregate demand that reduces the unemployment rate by 2 percentage points
- A negative shock that increases it by 2 percentage points.

1. What happens to the bargaining gap in each case?
2. What would you expect to happen to the price level in each case? Explain your answers.

14.4 INFLATION AND UNEMPLOYMENT: CONSTRAINTS AND PREFERENCES

Phillips' original curve, and the model in Figure 14.4d, suggest that there is a lasting trade-off between inflation and unemployment: for example, with the Phillips curve in the figure, if the government is happy to have inflation of 1% each year, then it can support a boom level of aggregate demand with an unemployment rate of 3%. If it prefers stable prices (zero inflation), then it needs to keep aggregate demand at the normal level, with unemployment of 6%. This suggests that the Phillips curve is a feasible set from which the policymaker can select the desired combination of unemployment and inflation.

Work through the slidebar in Figure 14.5 to see how the policymaker's preferences are described by indifference curves. To explain their shape, we focus on one of the indifference curves and note that:

- Where employment and inflation are high, the policymaker is happy to trade off a substantial drop in employment for a small reduction in inflation: this makes the indifference curve flatter.
- Where inflation and employment are low, the policymaker is happy to trade off a substantial increase in inflation for a small increase in employment: the indifference curves are steeper.
- The policymaker prefers the highest possible level of employment, shown by the labour supply. This means the indifference curves are horizontal when employment is equal to the labour supply.
- *X* marks the policymaker's preferred combination of inflation and unemployment.

As we saw at the end of section 14.1, the policymaker is likely to prefer low (stable) inflation to zero. This means the indifference curves become vertical at, say, 2% inflation. If inflation falls below this, the policymaker would be prepared to have lower employment if inflation could be raised.

Put another way: when the outcome is further from the inflation target but closer to full employment, the indifference curve is flatter because the policymaker places more value on getting closer to the inflation target. Conversely, when the outcome is further from full employment but closer to the inflation target, the indifference curve is steeper because the policymaker places more value on getting closer to full employment.

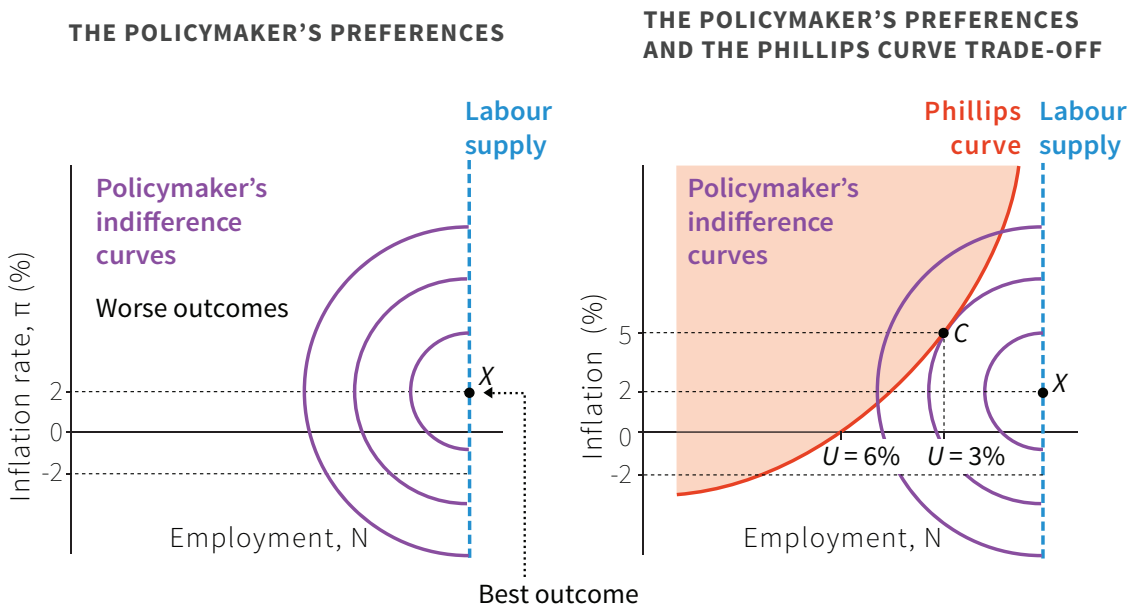


Figure 14.5 *The Phillips curve and the policymaker's preferences.*

In the right-hand panel of the figure, the indifference curves and the Phillips curve are shown. The policymaker sees the Phillips curve as the feasible set and will try to use monetary or fiscal policy to choose the level of aggregate demand so that employment is at *C*. This is the highest indifference curve consistent with the Phillips curve trade-off.

In this example, the policymaker prefers a combination of unemployment of 3% and inflation of 5% to the combination of unemployment of 6% and a stable price level (zero inflation).

DISCUSS 14.3: THE PHILLIPS CURVE AND THE POLICYMAKER'S PREFERENCES

The following questions refer to Figure 14.5.

1. What would the policymaker's indifference curves look like if the policymaker cared only about low unemployment?
2. Which point on the Phillips curve would that policymaker choose?
3. What would the policymaker's indifference curves look like if the policymaker cared only about low inflation?
4. Which point on the Phillips curve would this policymaker choose?
5. What would the indifference curves look like if the policymaker needed the support of pensioners (more than of working-age people) to be re-elected?

14.5 WHAT HAPPENED TO THE PHILLIPS CURVE?

The model in Figure 14.5 suggests that a policymaker who is able to adjust the level of aggregate demand can pick any combination of inflation and unemployment along the Phillips curve. But the data in Figure 14.6 suggests that the trade-off between inflation and unemployment is not a stable one. There is a mass of data points and no discernible positively sloped Phillips curve.

Figure 14.6 shows the inflation and unemployment combinations for the US for each year between 1960 and 2014. Note that on the horizontal axis the scale for the unemployment rate declines as we move to the right in the figure. A Phillips curve sketched through the observations in the 1960s gives a reasonably good picture of the inflation-unemployment trade-off in that decade. But that curve clearly does not fit in other periods.

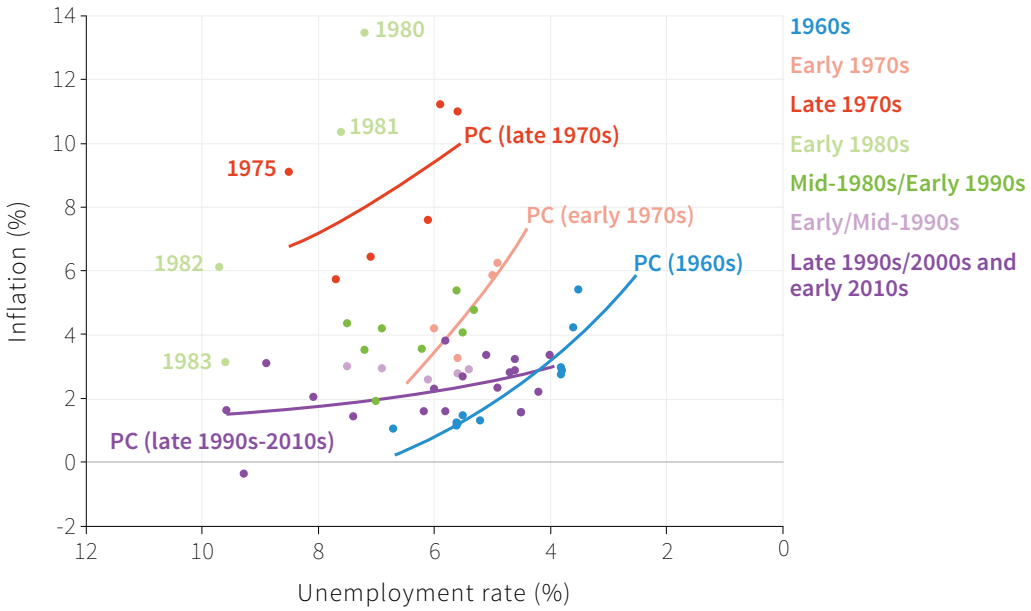


Figure 14.6 *Phillips curves in the United States (1960-2014).*

Source: *Federal Reserve Bank of St. Louis. 2015. 'FRED.' https://research.stlouisfed.org/fred2/*

We can use the figure to show how the Phillips curve shifted over time. The Phillips curve (PC) for the 1960s shows the economy was in a good state; the US could achieve combinations with relatively low inflation and unemployment. In this decade the dark blue dots show the individual years. Fluctuations of demand moved the economy along the Phillips curve but there were no shifts of the curve itself. In the early 1970s the Phillips curve appears to have shifted up. Each unemployment level was now associated with a higher level of inflation. The curve shifts again in the late 1970s; and again in the early 1980s, further worsening the trade off between

unemployment and inflation. In time, however, things improved: by the late 1990s, we can see that the Phillips curve had shifted down with lower unemployment and inflation.

In his presidential address to the American Economic Association in 1968, Milton Friedman provided an explanation for why the Phillips curve is not stable. He referred to the recent experience in the US. Between 1966 and 1968 unemployment had been steady, averaging 3.7%, but inflation had increased from 3.0% to 4.2%. He said that the only way unemployment could be kept as low as 3% was by allowing inflation to keep increasing: "There is always a temporary trade-off between inflation and unemployment; there is no permanent trade-off," he claimed. This is what Helmut Schmidt knew, but did not want to admit to the voters, in 1972.

If there is no permanent trade-off, then the Phillips curve is not a feasible set in the same way as the feasible consumption frontier was: the feasible consumption frontier stays in place when a different point on it is chosen. By contrast, Friedman, supported by evidence from many countries from the late 1960s, showed that:

If a government tries to keep unemployment "too low" the result will be not just higher inflation but rising inflation as well.

Inflation means rising prices. Rising inflation means prices increasing at an ever-faster rate. This means that the Phillips curve *would keep shifting upward*.

14.6 EXPECTED INFLATION AND THE PHILLIPS CURVE

We now explain why the Phillips curve shifts: why does inflation keep rising when governments try to keep unemployment too low? Why is there only one unemployment rate at which inflation is stable? We need to go back to two familiar points:

- *People are forward-looking:* We explained this in Units 6 and 12. They take actions now in anticipation of things they expect to happen. To stress this, economists say that "expectations matter".
- *People treat prices as messages:* As Friedrich Hayek taught us. Therefore they also treat changes in prices as messages about what will happen in the future, just as people treat a build-up of clouds as a prediction of rain.

With these two building blocks we can see why Friedman was right. As well as the battle for the pie between workers and the owners of firms that is the fundamental cause of rising *prices*, Friedman showed that, at low unemployment, inflation keeps

increasing. This is because of the way that wage- and price-setters form their views about what will happen to inflation, which is called *expected inflation*. The behaviour of inflation will reflect both elements.

Introducing expected inflation

We introduce the role of expected inflation by returning to the Phillips curve. Look at Figure 14.7. You will notice that, at labour market equilibrium with an unemployment rate of 6%, the inflation rate is 3% and not zero as in Figure 14.4d. If wage and price setters expect prices to rise by 3% per annum, and the level of aggregate demand is "normal", and keeps unemployment at 6%, then the economy can remain at the labour market equilibrium with inflation remaining constant at 3% per annum. Every year, wages and prices will rise by 3% and the real wage will remain at the intersection of the wage and profit curves. This is point A.

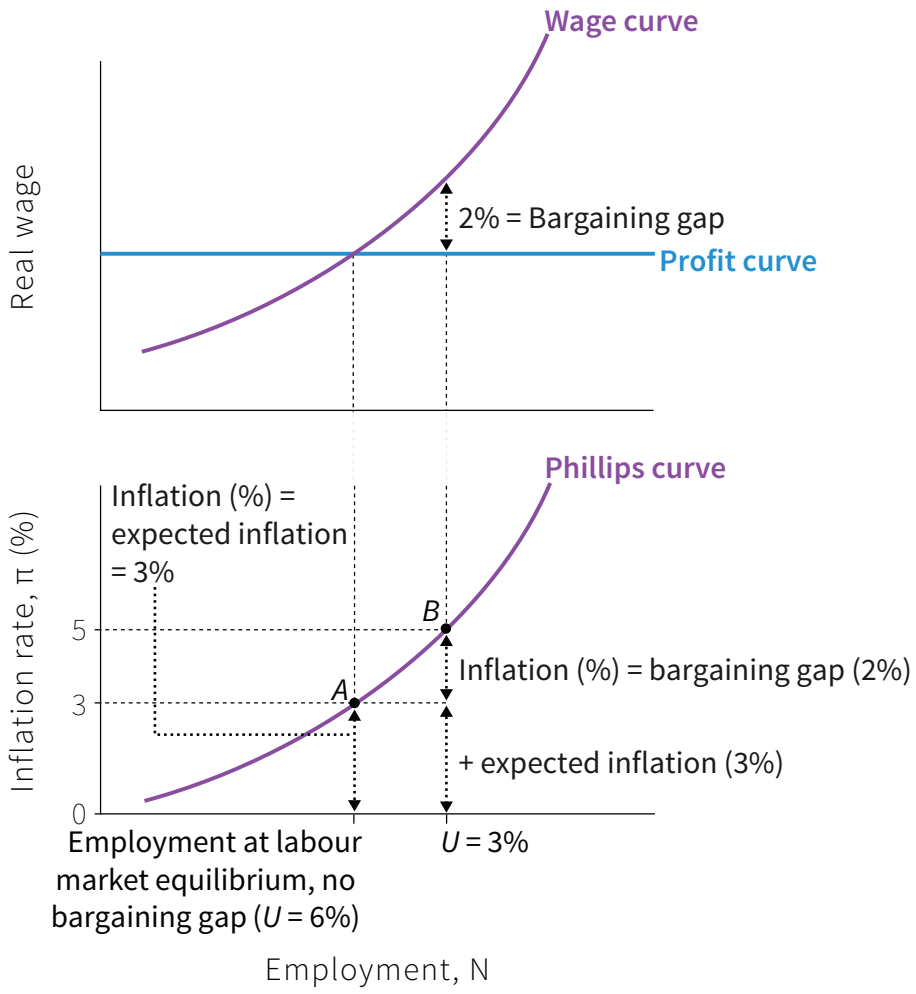


Figure 14.7 *Bargaining gaps, expected inflation and the Phillips curve.*

Now consider a boom, which takes the economy to lower unemployment at point B. What will happen to inflation? Workers expect prices to rise by 3% and will require a nominal wage increase of 3% just to keep their real wage unchanged. But they require an additional 2% rise to give them an expected real wage rise on the wage

curve, so wages increase by 5%. With their costs rising by 5%, firms will increase prices by 5%. In the boom, inflation will be 5%. This gives a Phillips curve like the one we have seen before: the only difference is that inflation at labour market equilibrium is 3% rather than zero.

When inflation is not zero, we can summarise the causal chain from the bargaining gap to inflation like this:

Expected inflation (%) + bargaining gap (%) → Increase in wages (%) → Increase in unit costs (%) (assuming no other costs) → Increase in prices (%) ≡ inflation (%)

To work out the inflation rate:

$$\begin{aligned} \textit{inflation (\%)} &\equiv \textit{increase in prices (\%)} \\ &= \textit{increase in costs per unit of output(\%)} \\ &= \textit{increase in wages (\%) (if wages are the only costs)} \\ &= \textit{expected inflation (\%) + bargaining gap (\%)} \end{aligned}$$

But Friedman pointed out that with low unemployment, inflation would not remain at 5% at point *B*. To see why, we ask what happens next?

The shifting Phillips curve

With low unemployment continuing, workers will be disappointed with the outcome, since they did not achieve their expected real wage. Why not? Workers expected a 2% real wage increase at *B* (to give the real wage on the wage curve) from their nominal pay rise of 5% but they did not get this because firms raised their prices by 5%.

But the story does not end there. We know that both parties cannot be satisfied with the outcome at low unemployment, because their claims add up to more than the size of the pie. At the next wage round, the human resources department has to take into account the fact that their employees expect prices to rise by 5%. This is based on the inflation experienced over the past year. Another interpretation is that HR includes inflation over the past year in the wage settlement, to make up for the shortfall in the real wage workers experienced because inflation turned out to be higher than expected. So in order to achieve another real wage increase of 2%, the HR department sets a wage increase of 7%. The process continues with the rate of inflation increasing over time.

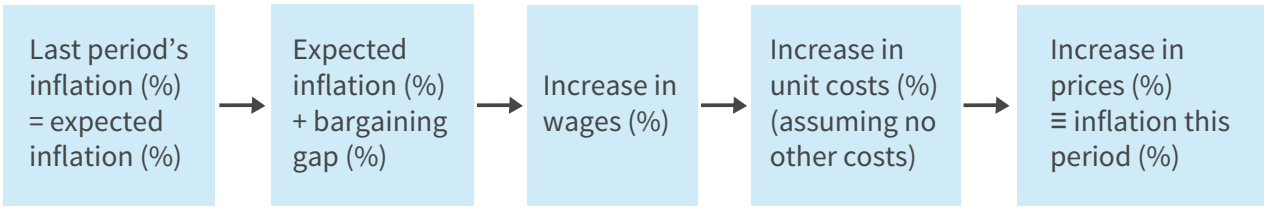
Figure 14.8 summarises the situation. We compare the situation over a three-year period with unemployment at two levels: 6% and 3%.

	Year	What inflation rate is expected? The inflation rate in the previous year (1)	Unemployment	Bargaining gap (3)	Inflation outcome: expectations plus bargaining gap (1)+(3)	Are wage setters and price setters disappointed (are their claims inconsistent)?
Stable inflation	1	3%	6%	0	3%	No
	2	3%	6%	0	3%	No
	3	3%	6%	0	3%	No
Rising inflation	1	3%	3%	2%	5%	Yes
	2	5%	3%	2%	7%	Yes
	3	7%	3%	2%	9%	Yes

Figure 14.8 *Unstable Phillips curves: Expected inflation and the bargaining gap.*

The first column of Figure 14.8 reflects forward-looking behaviour. Inflation expected over the year ahead is based on the previous year's inflation. The second column shows the unemployment rate. The third column shows the bargaining gap. The fourth column is the inflation outcome, which reflects expectations *and* the bargaining gap.

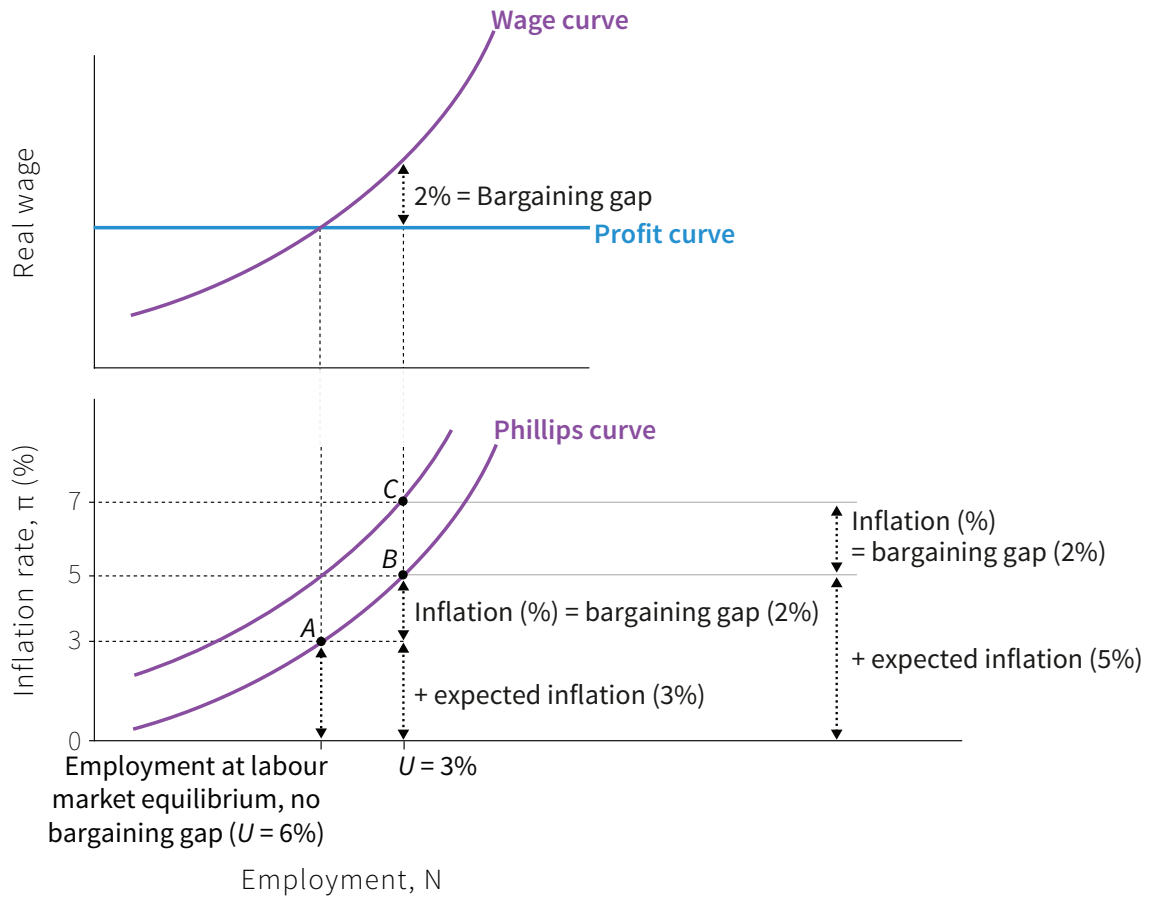
We can summarise the causal chain from the last period's inflation rate to this period's inflation rate like this:



To work out the inflation rate:

$$\begin{aligned} \textit{inflation (\%)} &\equiv \textit{increase in prices (\%)} \\ &= \textit{increase in costs per unit of output(\%)} \\ &= \textit{increase in wages (\%) (if wages are the only costs)} \\ &= \textit{expected inflation + bargaining gap (\%)} \\ &= \textit{last period's inflation + bargaining gap (\%)} \end{aligned}$$

We can show the data in Figure 14.8 in the Phillips curve and labour market diagrams. This is Figure 14.9. The stable inflation case is at point A with unemployment of 6% and inflation of 3%, year after year. At low unemployment (3%), the Phillips curve shifts up from the one through point *B* to the one through point *C* when expected inflation rises from 3% to 5%.



Labour market equilibrium

Inflation is 3% as expected.

A boom, first period

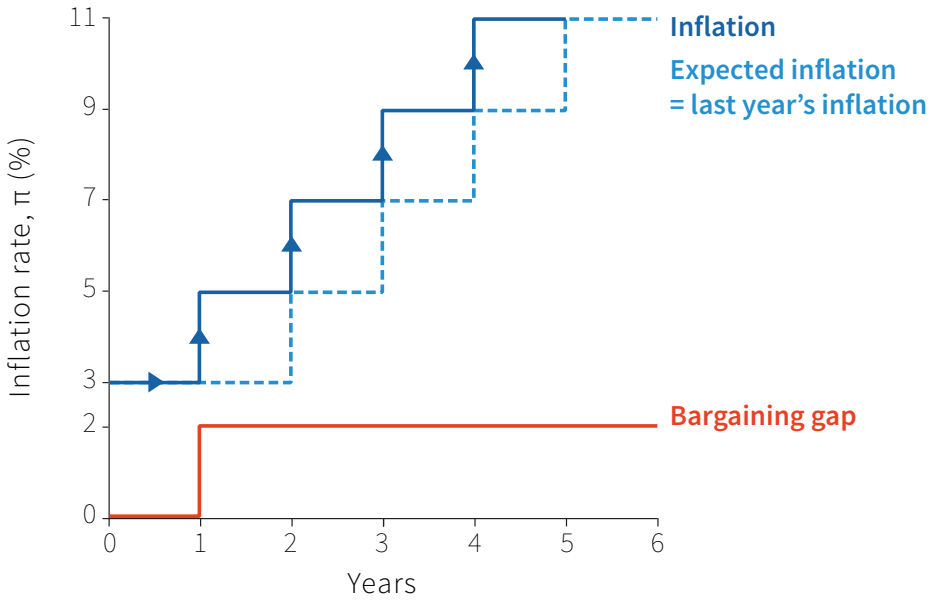
At lower unemployment, the bargaining gap is 2%. Inflation is equal to expected inflation plus the bargaining gap.

A boom, next period

Next period, with unemploy

Figure 14.9 *Inflation expectations and Phillips curves.*

By plotting the path of inflation over time in Figure 14.10 we can see the distinctive contributions to inflation of the bargaining gap and expected inflation. In this example, the bargaining gap opens up in year one because of the move to low unemployment. The assumption that unemployment remains below the *inflation-stabilising rate* is reflected in the persistence of the bargaining gap. Inflation rises every period because the previous period's inflation feeds into expected inflation and therefore into wage and price inflation. Note that the real wage does not change: the real wage is on the profit curve.



A zero bargaining gap

Inflation is as expected: 3%.

Year 1

At the start of year 1 following the opening up of the bargaining gap and after wages and prices have been adjusted, inflation is equal to the bargaining gap plus expected inflation.

Year 2

At the start of year 2, with no change in the bargaining gap, inflation goes up by the bargaining gap plus expected inflation.

... and each year afterwards

As long as the bargaining ga

Figure 14.10 *Inflation, expected inflation and the bargaining gap.*

DISCUSS 14.4: A NEGATIVE AGGREGATE DEMAND SHOCK WITH HIGH UNEMPLOYMENT

Copy Figure 14.9, making sure you leave plenty of space to the left of the 6% unemployment marker. Assume from an initial position at A, there is a negative shock to private sector demand such as depressed private investment, which raises unemployment to 9%.

1. Show the inflation, expected inflation and the bargaining gap at the new level of unemployment on your diagram.
2. What do you predict happens to inflation over the following two years, assuming there is no further change in unemployment?
3. Draw the Phillips curves and write a brief explanation of your findings.

DISCUSS 14.5: INFLATION, EXPECTED INFLATION AND THE BARGAINING GAP

Use the same axes as in Figure 14.10 to plot inflation, expected inflation and the bargaining gap in the following cases:

1. The price level is constant in period zero in the economy shown in Figure 14.10.
2. The economy of Figure 14.10 is hit by a recession at the beginning of period 1 and unemployment remains at a constant high level over the subsequent years.
3. At the beginning of period 6, the bargaining gap disappears. Give a brief explanation of why this might have happened and any other assumptions you are making.
4. The economy is represented by the Phillips curve model presented in Figure 14.7 in which the Phillips curve does not shift due to inflation expectations.

14.7 SUPPLY SHOCKS AND INFLATION

Friedman was correct in two ways:

- Expected inflation shifts the Phillips curve.
- Policymakers were wrong to think of the Phillips curve as a feasible set from which they could simply select the most electorally popular combination of inflation and unemployment.

But there are other causes of high and rising inflation. The Phillips curve will shift up if the profit curve shifts down or the wage curve shifts up. Recall Figure 14.2: if the power of owners of firms relative to consumers increases, the marketing department raises prices and kicks off a wage-price spiral. In that example, owners of firms in the home economy became more powerful because the government adopted policies to make it more difficult for foreign firms to enter the economy. Similarly, a wage-price spiral can begin if the power of employees increases relative to owners—as would be the case if trade unions become more powerful and exercise that power to achieve higher wage increases from the HR department.

Changes in the global economy can also trigger inflation. A particularly important change for understanding the shifts in Phillips curves, such as those for the US economy shown in Figure 14.6, is a change in the world oil price. (We look at other possible causes in Units 15 and 17.) The labour market model and the Phillips curve can explain why a one-off increase in the world oil price can lead to:

- A rising price level (inflation)
- Rising inflation

To do this, we show that a rise in the oil price:

- *Shifts the profit curve down:* This leads to a positive bargaining gap and inflation.
- *Shifts the Phillips curve up:* It will continue to shift up as expected inflation rises.

An increase in the oil price pushes down the profit curve. A typical firm uses imported oil in the production process. With increased costs for oil, the firm's profits can only remain unchanged if real wages fall. At the level of the economy as a whole, the national pie to be divided between owners and employees shrinks when more has to be paid for imports.

We show in the Einstein section how to modify the profit curve once firms in the economy use imported materials in production.

The mechanism through which an oil price hike creates a bargaining gap and triggers a wage price spiral is through its effect on the price level. Firms raise their prices to protect their profit margins when the cost of imported oil rises. This reduces the real wage of employees because the price level in the economy rises (to see how firms set their prices following an oil price rise, see the Einstein section).

Remember that the profit curve shows the real wage after firms have set their prices. Given the competitive conditions in the economy, the firm's marketing department will pass on the higher energy costs per unit of output. Firms across the economy will behave this way and the price level will rise. This downward shift in the profit curve opens up a bargaining gap between the real wage that has to be paid to get workers to work effectively, and the real wage on the profit curve.

Following the oil price rise, the profit curve in Figure 14.11 shifts down and, in this example, a bargaining gap of 2% opens up between the wage curve and the post-shock profit curve. This fits the scenario in Figure 14.10, where a bargaining gap of 2% appears at the beginning of year 1. This increases inflation from its pre-existing level of 3% to 5% and as expected inflation adjusts, inflation rises thereafter every year. The Phillips curve shifts up year by year.

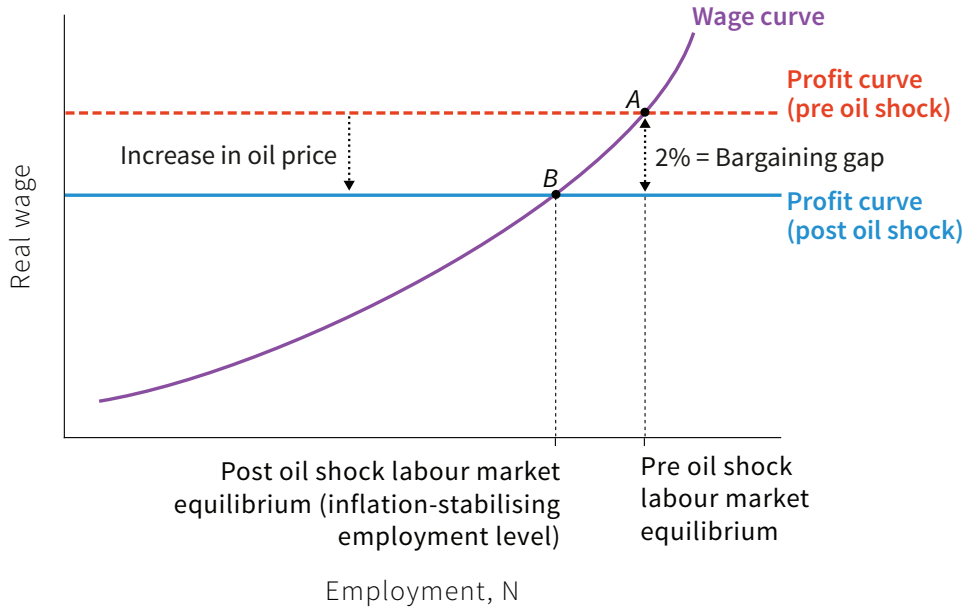


Figure 14.11 *An oil shock and the profit curve.*

As long as employment remains at its pre-oil shock level, inflation will increase every period, as illustrated in Figure 14.10. The new post-shock inflation-stabilising employment rate is shown in Figure 14.11. Unemployment is higher at the new labour market equilibrium where the post-shock profit curve intersects the wage curve.

Shocks to the world oil price are a major source of macroeconomic disturbance. Following the early 1970s oil shock, for example, US inflation jumped from 6.2% in 1973 to 9.1% in 1975 and unemployment went from 4.9% to 8.5% at the same time.

This pattern was common across the developed world. For example, in the same period, inflation in Spain rose from 11.4% to 17% and unemployment increased from 2.7% to 4.7%.

We can see from Figure 14.12 that there were two big recessions in the UK in the 1970s. They were due to the oil shocks of 1973-74 and 1979-80, which were associated with a rise in both unemployment and inflation to post-second world war peaks (you can see the effect on inflation in Figure 12.19b and Figure 12.19c).

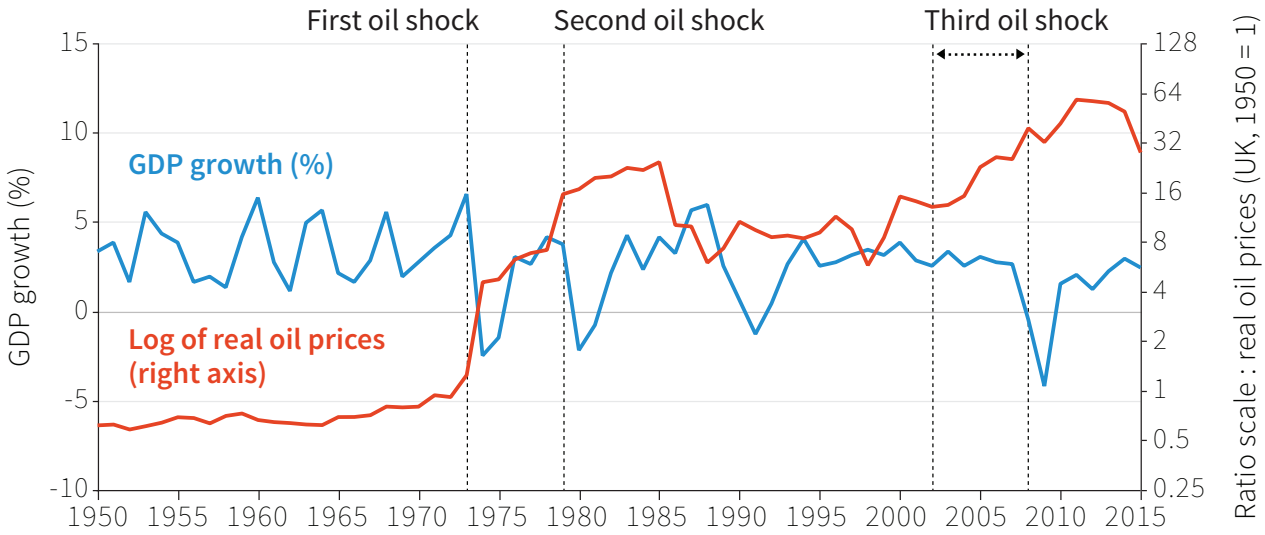


Figure 14.12 *UK GDP growth and real oil prices (1950-2015).*

Source: *UK Office for National Statistics; "Three centuries of Macroeconomics data", Bank of England*

High inflation in the 1970s and early 1980s was associated with high unemployment in many countries. Unemployment in the UK peaked at nearly 12% in the mid-1980s. The model helps us to understand the role such high unemployment played in bringing inflation down.

In the model, the only ways that high inflation can be brought down are:

- A reduction in the bargaining gap
- A fall in expected inflation

If unemployment is sufficiently high, then there will be a negative bargaining gap and inflation will fall. Remember that for the bargaining gap to be negative, unemployment has to rise above the new higher inflation-stabilising unemployment rate. Once inflation begins to fall, it will continue to fall as the Phillips curve shifts downwards and the economy follows the path shown in Figure 14.10 in reverse.

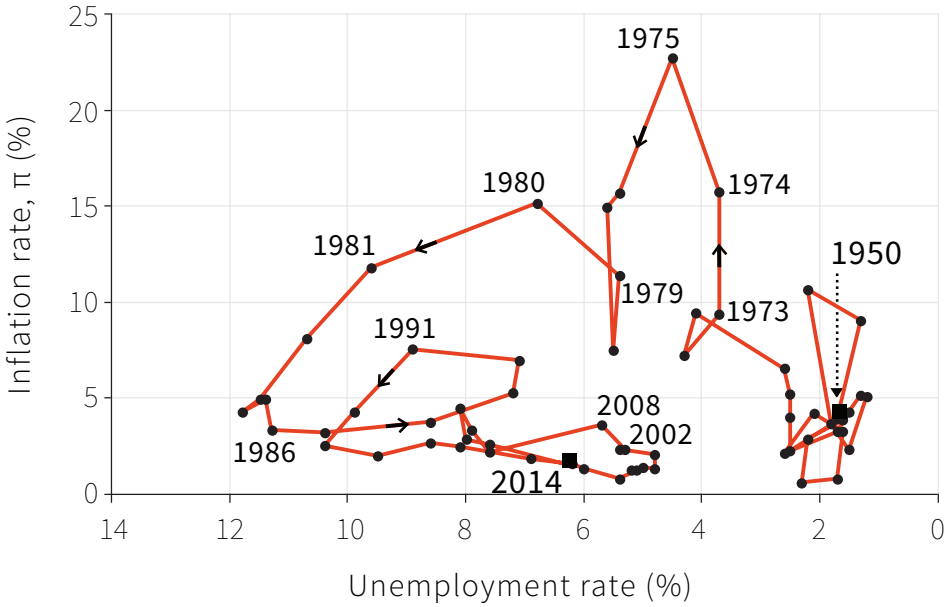


Figure 14.13 *UK Inflation and unemployment rate (1950-2014).*

Source: *UK Office for National Statistics; "Three centuries of Macroeconomics data", Bank of England.*

Figure 14.13 shows a scatterplot of unemployment and inflation for the British economy from 1950 to 2014. Instead of fitting Phillips curves to the observations, as in Figure 14.6, the points are joined and dated. This helps us to follow the path taken by the economy. Notice the large increase in unemployment in the 1980s associated with bringing inflation down: this is sometimes referred to as the *cost of disinflation*.

But there's a puzzle here: why did the third oil shock from 2002-2008 not lead to increased inflation, just like the earlier ones? This section should have provided you with some starting points to investigate this, and this speech, given by an economist in 2006, will help you. If you read both carefully, you might ask the following questions:

- *Was the unit cost increase smaller due to less energy-intensive production*? This would have made the increase in the materials cost per unit of output smaller and reduced the size of the initial downward shift in the profit curve.
- *Did the wage curve shift downwards* at the same time as the third oil price shock? This also would have reduced or perhaps even eliminated the bargaining gap opened up by the oil price shock.
- *Did a wage-price spiral fail to develop* because expected inflation did not adjust upward, as in the past oil shocks?

DISCUSS 14.6: AN OIL SHOCK

Think about the three questions we listed above. In each case:

1. Explain the mechanism linking the oil shock to inflation using a diagram.
2. Identify some evidence (for example, data or commentary in the economics press) that is consistent with the hypothesis proposed.

What could stop *expected* inflation rising? Perhaps changes in monetary policy? We examine this in the next section.

14.8 MONETARY POLICY

We use the Phillips curve and the policymaker's indifference curves to look at shocks and policy responses. Before doing so, we need to recall how monetary policy affects the economy.

As we saw, we can explain why people might dislike rising inflation, but not why they should dislike a (slowly) rising price level. In fact, many central banks around the world have policies to target an inflation rate of 2%. They either set this objective for themselves, or the government sets the objective for them. It means they are doing best if prices rise each year by a rate close to 2%.

When central banks target an inflation rate of 2%, the best answer to the question "why does the price level rise?" becomes "because the central bank makes it happen". As we first saw in Unit 11, when inflation is forecast to be higher or lower than this, the central bank can take action to adjust the level of aggregate demand and employment so as to steer the economy toward a 2% target.

When they can, central banks use changes in the policy interest rate as their monetary policy instrument to stabilise the economy. Monetary policy relies on the central bank being able to control interest rates, and on changes in interest rates influencing aggregate demand. For example, higher interest rates make it more expensive to borrow money to spend. It is important to remember that it is the real interest rate that affects spending. But when the central bank sets the policy rate,

it sets it in nominal terms. So by setting a particular nominal rate it is aiming for a specific real interest rate, and it therefore takes account of the effect of expected inflation (see our Einstein *The real interest rate* on the Fisher equation).

Figure 14.14 shows how the Bank of England views the transmission of monetary policy from its interest rate decision to aggregate demand and inflation in "normal" situations—that is, when the interest rate is its policy instrument.

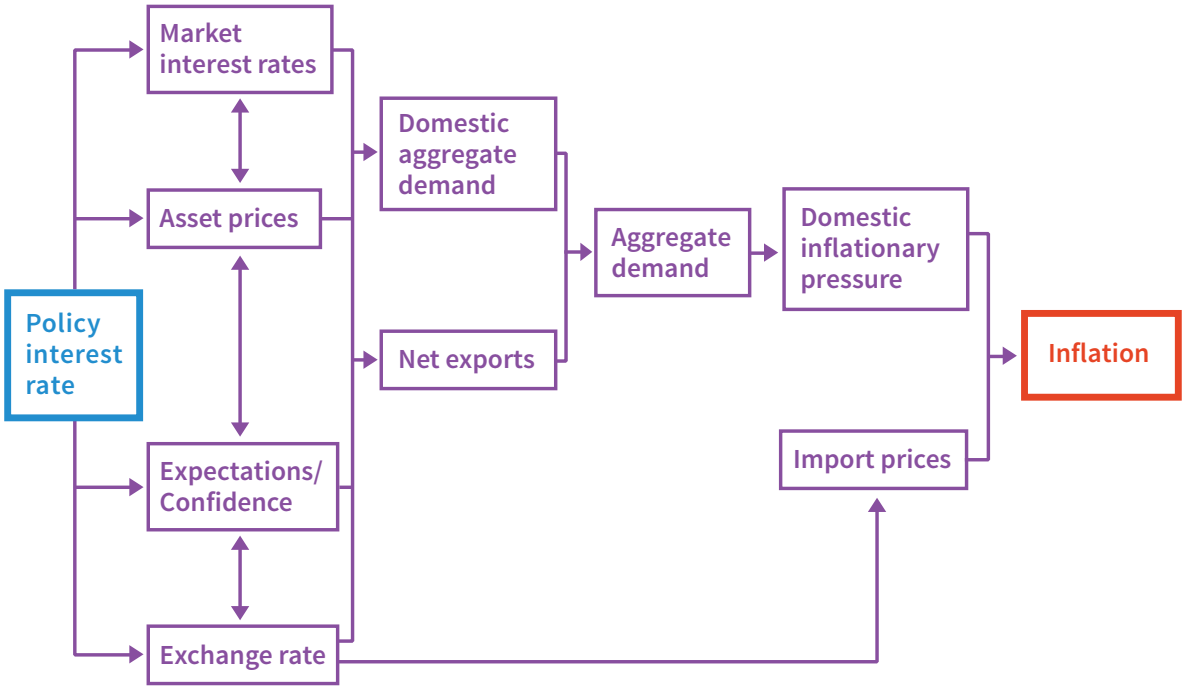


Figure 14.14 *Monetary policy transmission mechanisms.*

Source: *Bank of England, 2012.*

Look at the boxes in the first column:

Market interest rates

In Unit 13 we explained that, although the central bank sets the *policy interest rate*, commercial banks set the *market rates* that households and firms pay when they take out loans. When the central bank cuts the policy rate to stimulate spending, the market interest rate will fall by approximately the same amount. To set the policy rate, the central bank will therefore work backwards, starting with its desired level of aggregate demand:

1. It will estimate a target for the total aggregate demand, Y, to stabilise the economy, and the *real interest rate*, r, that will produce it (this shifts the aggregate demand line into the desired position in the multiplier diagram).
2. It calculates the market interest rate that will create that level of aggregate demand.

3. Finally it calculates the nominal policy rate, i, that will produce the appropriate market interest rate.

Think about how a fall in the market interest rate affects the decision to build a new house. The cost of taking out a loan to finance the construction of the house will fall; so, as the interest rate falls, investors will consider more new housing projects to be financially viable. Through this channel a lower policy rate will raise investment—by businesses and households—and a higher policy rate will lower it (see Figure 13.9).

Asset prices

This refers to financial assets in the economy such as government bonds and shares issued by companies. When the central bank changes the interest rate, this has a ripple effect through all the interest rates in the economy—from mortgage rates to the interest rates on 20-year government bonds. When the interest rate goes down, the price of the asset goes up. So a fall in interest rates will be expected to feed through to spending, because households who own the assets will feel wealthier.

Profit expectations and confidence

In Unit 12 and Unit 13 we stressed the importance of profit expectations and confidence for the investment decisions of firms. When setting the interest rate, the central bank tries to build confidence through consistent policymaking and good communications with the public. If it lowers the policy rate and explains its reasoning, this can lead firms and households to bring forward investment projects.

Exchange rate

We return in the next section to the way monetary policy affects aggregate demand through the exchange rate channel: this will shift the aggregate demand line by changing net exports, $(X - M)$.

In the multiplier model of aggregate demand, the transmission channels from the policy rate to domestic aggregate demand are reflected in the investment function (including new housing), which shifts when the real interest rate changes: we write this function $I(r)$. The expectations and asset price effects will shift the investment function as we saw in Figure 13.5, and the consumption function, by changing c_0 (Figure 13.11).

In the multiplier diagram, the intercept of the aggregate demand line with the vertical axis includes investment, which means that the line shifts whenever the interest rate is changed by the central bank, or when business confidence changes. If the central bank is trying to boost the economy in a business cycle downturn, it cuts the interest rate. By signalling its willingness to support growth, the central bank also aims to influence the confidence of decision-makers in firms and households and help shift the economy from the low-investment equilibrium illustrated in the coordination game in Unit 12 to a high-investment equilibrium.

Figure 14.15 shows how monetary policy can be employed to stabilise the economy following a downturn caused by a drop in consumption (for example, as a result of a fall in consumer confidence). The economy starts in goods market equilibrium at point *A*. Consumption then falls from c_0 to c_0', which shifts the aggregate demand line down and the economy enters recession, moving from point *A* to point *B*. In order to stabilise the economy, the central bank stimulates investment by lowering the real interest rate from *r* to *r'*. This policy shifts the aggregate demand curve upward, pulling the economy out of recession and back to its starting point.

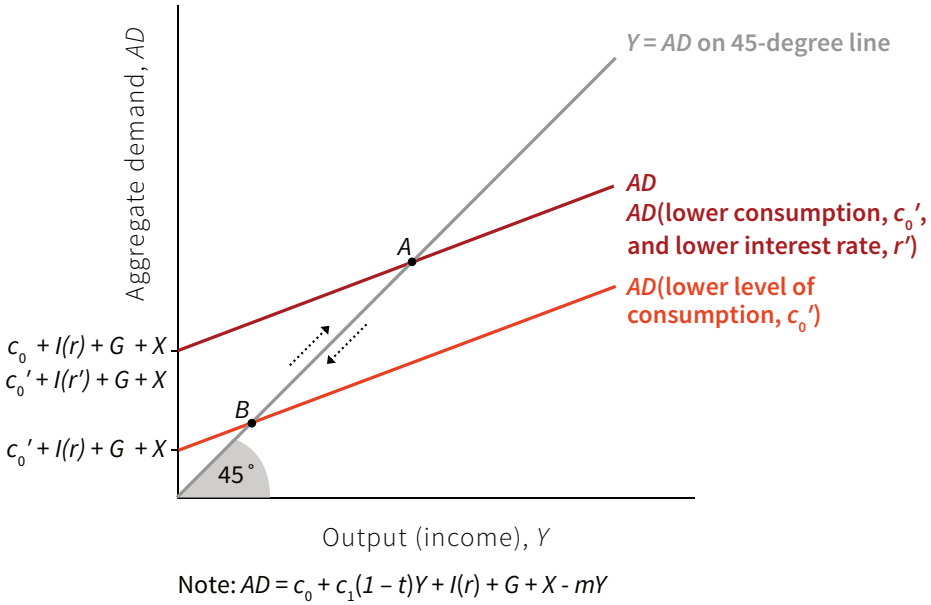


Figure 14.15 *The use of monetary policy to stabilise the economy in a recession.*

A warning

Using simple diagrams like Figure 14.15 may give the impression that the central bank is able to stabilise the economy by accurate diagnosis of a shock and precise intervention with a change in the interest rate. This is far from the case! The economy emits all kinds of noisy signals and it is difficult to decide, for example, whether a downturn is a temporary blip or signifies long-term weakness. The models we use help us to organise our thinking about the causal links in the economy and what policies might be warranted. *They do not give a complete recipe for effective stabilisation*.

Figure 14.15 shows how the central bank could attempt to counteract a recession. But how should the central bank react to a consumption boom? It needs the opposite policy. A boom will shift the aggregate demand line upwards, so the central bank must pursue policies that dampen demand and return the aggregate demand line back to its starting point. The central bank can do this by raising the interest rate.

But why would it want to curtail a boom? From the Phillips curve, we know that a boom leads to higher inflation, and, if expectations adjust to past inflation, to rising inflation. High and rising inflation imposes costs on the economy.

We have shown how monetary policy can be used by the central bank to stabilise the economy in a recession. The government could also have played this role by cutting taxes, or by boosting spending.

Why monetary policy, and what are its limits? Fiscal policy is complicated to adjust and inflexible. Instead, to keep aggregate demand close to the level it desires, the central bank can adjust the interest rate up and down by small amounts month-by-month.

However, there are two important limitations to the usefulness of monetary policy in stabilisation:

- *The short-term nominal interest rate cannot go below zero:* But this is the central bank's policy instrument.
- *A country without its own currency does not have its own monetary policy.*

The zero lower bound

If the policy interest rate were negative, people would simply hold cash rather than put it in the bank, because they would have to pay the bank for holding their money (that's what a negative interest rate means). This is the *zero lower bound* on the nominal interest rate. It matters because, when the economy is in a slump, a nominal interest rate of zero may not be low enough to achieve a sufficiently low real interest rate to drive up interest-sensitive spending and get the economy going again. This is the reason why economies that were badly hit by the global financial crisis introduced a new kind of monetary policy called *quantitative easing (QE)*. The aim of QE is to affect boost aggregate demand by buying assets.

How is QE supposed to work?

- *The central bank buys bonds and other financial assets:* More demand for these assets pushes up their price (remember that the prices of financial assets will change when the demand curve or supply curve shifts as we saw in Unit 9).
- *This creates demand for financial assets:* So the central bank shifts the demand curve to the right, which pushes up the price.
- *The price of bonds rises:* The yield or interest rate on the assets falls. Reread the *glossary definition* to remind yourself why this is true.
- *This boosts spending:* Particularly on housing and consumer durables, because both the cost of borrowing and return to holding financial assets has gone down.

So, even when the interest rate the central bank directly controls is stuck at zero, it can use QE to try to reduce the interest rate on a variety of other financial assets. The empirical evidence suggests that the effects of QE in boosting aggregate demand are positive but small.

No national monetary policy

Monetary policy may not be available to a country. Members of the eurozone gave up their own monetary policy when they joined the currency union. The eurozone is called a *common currency area* because all the members use the euro. This means there is just one monetary policy for the whole of the eurozone. The European Central Bank in Frankfurt sets the policy interest rate, because it controls the base money used by all banks in the eurozone. This interest rate may be more appropriate for some members than for others. In particular, after the financial crisis, unemployment was low and falling in Germany but, in the southern eurozone countries such as Spain and Greece, it was high and rising fast. There were many complaints that the ECB's monetary policy remained too restrictive for too long for the needs of the latter countries.

DISCUSS 14.7: FISCAL OR MONETARY POLICY?

Think back to the discussion of the government finances in Unit 13.

1. In the event of a financial crisis, would it be preferable for the government to stabilise the economy using fiscal or monetary policy?
2. What are the dangers of using fiscal policy?
3. When might the government have no choice but to use fiscal policy?

Once we open our eyes to the fact that the economy is embedded in a global economy, we see there is another way for monetary policy to affect aggregate demand—through the *exchange rate*.

14.9 THE EXCHANGE RATE CHANNEL OF MONETARY POLICY

Monetary policy in the US works mainly through the effect of changes in the interest rate on investment, particularly on new housing, and on consumer durables. But, in many other economies, especially smaller ones, an important channel for monetary policy is through the effect of interest rate changes on the *exchange rate* and the economy's competitiveness in international markets, and hence on net exports.

Take the case of a slowdown in the Australian economy. The Reserve Bank of Australia responds to this by cutting the interest rate. The cut in the interest rate leads to a depreciation of the Australian dollar, which means that it will buy a smaller number of US dollars, Chinese yuan, euros, or any other currency. Depreciation makes Australian exports and home-produced goods more competitive, boosting aggregate demand and stabilising the economy. Both higher export demand for home's products (X) and lower demand by home for goods and services produced abroad (M) raise aggregate demand in the home economy.

But there is a missing link in this argument: how does the cut in the interest rate by the Reserve Bank lead to a depreciation of the Australian dollar? Home's exchange rate depreciates because of a fall in the demand for Australian dollars in the foreign exchange market. Why does a cut in the interest rate have this effect? Australian financial assets, on which interest is earned, become less attractive to international investors when the interest rate is cut. For example, when the Reserve Bank of Australia reduces the interest rate, there is less demand for a typical financial asset sold by the Australian government, such as a 3-month or 10-year government bond. If the demand for Australian financial assets like government bonds goes down, then the demand for the Australian dollars needed to buy them also goes down.

The foreign exchange market is a market in which currencies are traded against each other, in this case the Australian dollar (AUD) and the US dollar (USD). The exchange rate is defined as the number of units of home currency for one unit of foreign currency, in other words:

$$\textit{exchange rate of the Australian dollar} = \frac{\textit{number of AUD}}{\textit{one USD}}$$

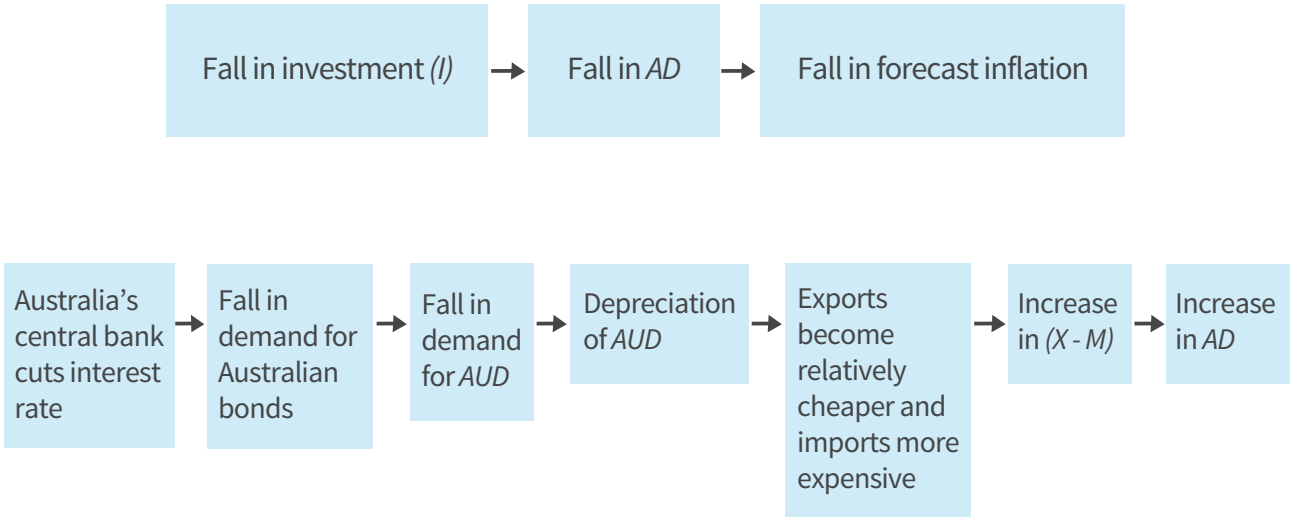
When one USD buys more AUD, the AUD is said to have *depreciated*. When one AUD buys more USD, the AUD is said to have *appreciated*.

A depreciation of home's exchange rate makes home's exports cheaper, and imports from abroad more expensive. For example, if a t-shirt in Australia costs 20 AUD, and the exchange rate with the USD is 1.07 (remember this is the number of AUD/one USD), then the t-shirt costs $20/1.07 = 18.69$ USD in the US. Equivalently, a t-shirt sold for 18.69 USD in the US would cost 20 AUD in Australia. If the exchange rate of

the Australian dollar then depreciates to 1.25, what happens to the price of exports and imports of t-shirts in Australia? Exports of Australian t-shirts become cheaper; a 20 AUD t-shirt now costs only 16 USD in the US rather than 18.69 USD. In contrast, imports of US t-shirts into Australia become more expensive, a 18.69 USD t-shirt now costs 23.36 AUD rather than 20 AUD.

When the home exchange rate depreciates, the home economy's products become relatively cheaper on the world market. This increases the demand coming from foreigners for home's goods and services, which means that exports go up. Since the depreciation also makes the purchase of goods and services from abroad less attractive because they are now relatively more expensive, imports go down. The combination of higher exports and lower imports increases aggregate demand for domestic output and helps offset the initial shock that provoked the interest rate cut.

A rough summary of the chain of events in Australia goes as follows:



DISCUSS 14.8: WHY BONDS?

Explain why a change in the central bank's policy interest rate affects the exchange rate through the market for financial assets (such as government bonds), rather than through the market for goods and services.

14.10 DEMAND SHOCKS AND DEMAND-SIDE POLICIES

To see how policymakers respond to demand shocks in practice, think about the recession in the US after the bursting of the tech bubble. Figure 14.16 illustrates the fiscal and monetary policy mix used during the US recession in 2001 when, after a decade of expansion, the growth rate of the US economy slowed. The top row shows that the annual growth rate of real GDP decreased from 4.1% to 0.9%. The bottom two rows in Figure 14.16 show that the slowdown led to rising unemployment and falling inflation, exactly as we would expect from a negative demand shock. The end of the boom of the late 1990s, during which firms had been overoptimistic about the profits to be made on investment in new technology and had overestimated the need for new capacity in ICT-producing industries, triggered the slowdown (see Unit 9 for more about the tech bubble and Figure 13.5 for the model of investment with supply and demand effects shifting the investment function).

		2000	2001	2002	2003
Real Gross Domestic Product (annual%change)		4.1	0.9	1.8	2.8
Contribution to % change in GDP	Change in nonresidential investment	1.15	-1.2	-0.66	0.69
	Change in residential investment	-0.07	0.09	0.39	0.66
	Change in government expenditure	0.10	0.88	0.74	0.36
	Change in other contributions	2.92	1.13	1.33	1.09
Federal Reserve nominal interest rate (annual average, %)		6.24	3.89	1.67	1.13
Unemployment rate (%)		4	4.7	5.8	6
Inflation rate (%)		3.4	2.8	1.6	2.3

Figure 14.16 *The policy mix: Fiscal and monetary policy in the US following the collapse of the tech bubble.*

Source: *Federal Reserve Bank of St. Louis. 2015. 'FRED.'*

The recession and the policy response

The figure shows that the contribution of nonresidential investment to the percentage change in GDP was much larger than either residential investment or government expenditure in 2000. It fell in 2001, pulling the economy into recession. The recession could have been much worse in the absence of the strong response from monetary and fiscal policy.

In 2001, the Federal Reserve started rapidly decreasing the nominal interest rate, from a high of 6.2% on average in 2000, to 3.9% in 2001, and a low of 1.1% in 2003.

- *Monetary policy:* We can see from Figure 14.16 that this large drop in nominal interest rates helped boost residential investment in 2001 and 2002. Its contribution to growth became much larger than before. It also helped nonresidential investment to recover, but the adjustment was slower: the contribution of nonresidential investment to growth became positive only in 2003.
- *Fiscal policy:* To compensate for the stagnation in firms' private investment, the government used an expansionary fiscal policy. It introduced large tax cuts and an increase in spending in 2001 and 2002. The multiplier model helps explain the logic of the government's policy, and the large increase in the contribution of public expenditure to growth in 2001 and 2002.

We can see from Figure 14.16 that the swift action of the government and central bank helped to stabilise the economy; inflation and GDP growth bounced back rapidly from the recession. Unemployment was slower to react, however, continuing to creep up in 2003. In fact, the US unemployment rate has never dropped all the way down to its 2000 level, perhaps suggesting that the US economy was operating above capacity in the run-up to the tech bubble.

The recession and the model

We can apply the model we have developed to the case of a slump in investment spending in the US economy in Figure 14.16. From the multiplier diagram in the lower panel, we know that a fall in investment spending shifts the aggregate demand line down, and leads to a new equilibrium in the economy with lower output and higher unemployment. Figure 14.16 showed that this is what happened in the US after the tech bubble ended. Unemployment increased from 4% in 2000 to 6% in 2003, and inflation fell from 3.4% in 2000 to 1.6% in 2002.

Following the logic of the Phillips curve, inflation will fall in response to a rise in unemployment. Work through the slidebar in Figure 14.17 to see the consequences of the shock. As we can see in the left-hand panel of Figure 14.17, the economy moves to a situation with higher unemployment and lower inflation (from point *C* to point *D*). It is clear from the indifference curves that wellbeing in the economy has gone down. Remember that the best outcome is not full employment: it is the level of

employment (and unemployment) that maintains labour market equilibrium, to avoid consistently rising inflation. In Figure 14.17, point *X* is the policymaker's best outcome: inflation is at target and employment is consistent with constant inflation.

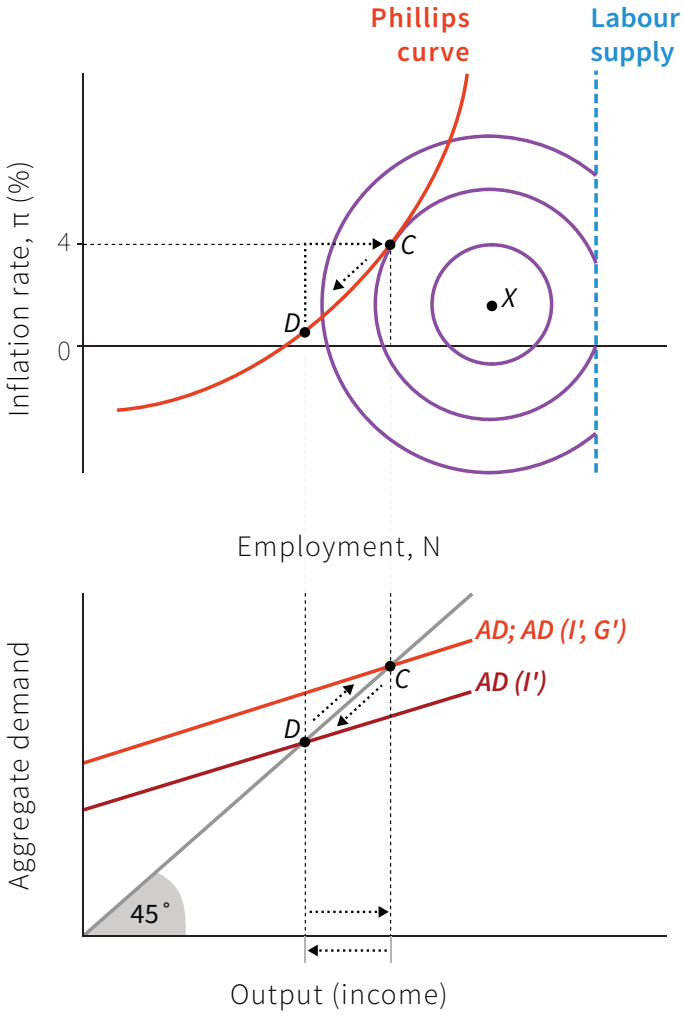


Figure 14.17 *Policy intervention to restore employment and output after a fall in investment.*

What can the policymaker do to restore wellbeing? We can see from the multiplier diagram in the bottom panel of Figure 14.17 that the fall in investment shifts the aggregate demand line down, reducing output. The government can combat the investment slump by increasing its spending, which shifts the aggregate demand line back to its starting position. Public spending replaces the lost private spending, and output is restored. The increase in output from the increased government spending reduces unemployment and raises inflation; the economy moves back along the Phillips curve to point *C*. The policy intervention has restored output and employment in the economy. We have shown the intervention as an increase in government spending, but demand could equally have been stimulated by a reduction in the tax rate or the interest rate. As we saw in Figure 14.16, both fiscal policy measures—tax cuts and spending increases—and monetary policy were used in the US in 2001.

DISCUSS 14.9: A CONSTRUCTION BOOM

1. What happens when there is a positive shock to aggregate demand from a boom in the construction of new housing? Explain using the multiplier diagram and the Phillips curve diagram.
2. What would you expect the central bank to do?

14.11 MACROECONOMIC POLICY BEFORE THE GLOBAL FINANCIAL CRISIS: INFLATION-TARGETING POLICY

The 15 years before the global financial crisis in 2008 came to be known as the *great moderation*. A look back at Figure 14.12 tells us why. Despite a major oil shock in the 2000s, the British economy and many other economies continued to experience steady growth, low inflation and low unemployment. This is a remarkable contrast with the high inflation and high unemployment of the 1970s.

There were two important features of the 1990s and 2000s prior to the crisis:

- *Central banks were made independent of government control:* Monetary policy was placed in the hands of these independent central banks in most advanced and many developing countries.
- *Inflation targeting:* These banks use their policy instruments to keep the economy close to a target rate of inflation. As shown in Figure 14.18, by 2012 28 countries had adopted inflation targeting, usually with a band (range) of what was judged an acceptable level of inflation.

Why make central banks independent and give them inflation targets? The lessons of Figure 14.6 about the instability of Phillips curves, and the high costs of unemployment incurred by countries in the 1980s as they brought inflation down, created the impetus. Policymakers globally believed there would be an inflation-stabilising unemployment rate.

Beginning in the 1990s, governments increasingly took the view that central banks should be given responsibility for keeping the economy close to a target rate of inflation. This is typically around 2% in developed economies, but higher in some developing economies, as Figure 14.18 shows. Since many voters will prefer lower

unemployment even if it comes with higher inflation, as we saw in section 14.1, how can central banks credibly commit not to deviate from their announced inflation target?

COUNTRY	INFLATION TARGETING ADOPTION DATE	INFLATION RATE AT ADOPTION DATE (%)	2010 END-OF-YEAR INFLATION (%)	TARGET INFLATION RATE (%)
NEW ZEALAND	1990	3.30	4.03	1 - 3
CANADA	1991	6.90	2.23	2 +/- 1
UNITED KINGDOM	1992	4.00	3.39	2
AUSTRALIA	1993	2.00	2.65	2 - 3
SWEDEN	1993	1.80	2.10	2
CZECH REPUBLIC	1997	6.80	2.00	3 +/- 1
ISRAEL	1997	8.10	2.62	2 +/- 1
POLAND	1998	10.60	3.10	2.5 +/- 1
BRAZIL	1999	3.30	5.91	4.5 +/- 1
CHILE	1999	3.20	2.97	3 +/- 1
COLOMBIA	1999	9.30	3.17	2 - 4
SOUTH AFRICA	2000	2.60	3.50	3 - 6
THAILAND	2000	0.80	3.05	0.5 - 3
HUNGARY	2001	10.80	4.20	3 +/- 1
MEXICO	2001	9.00	4.40	3 +/- 1
ICELAND	2001	4.10	2.37	2.5 +/- 1.5
KOREA, REPUBLIC OF	2001	2.90	3.51	3 +/- 1
NORWAY	2001	3.60	2.76	2.5 +/- 1
PERU	2002	-0.10	2.08	2 +/- 1
PHILIPPINES	2002	4.50	3.00	4 +/- 1
GUATEMALA	2005	9.20	5.39	5 +/- 1
INDONESIA	2005	7.40	6.96	5 +/- 1
ROMANIA	2005	9.30	8.00	3 +/- 1
SERBIA	2006	10.80	10.29	4 - 8
TURKEY	2006	7.70	6.40	5.5 +/- 2
ARMENIA	2006	5.20	9.35	4.5 +/- 1.5
GHANA	2007	10.50	8.58	8.5 +/- 2
ALBANIA	2009	3.70	3.40	3 +/- 1

Figure 14.18 *Countries with inflation-targeting central banks by 2012.*

Source: *Jahan, Sarwat. 2012. 'Inflation Targeting: Holding the Line.' International Monetary Fund Finance & Development.*

To tackle this concern, many countries increased the independence of the central bank. Politicians, like the West German superminister Helmut Schmidt, may want to promise lower unemployment now—even if this leads to inflation later—to be re-elected. Making the central bank independent, with an explicit inflation target, makes it easier for the central bank to resist political pressure. This prevents an inflation spiral.

Figure 14.19 illustrates the relation between the degree of central bank independence in the mid-1980s, and average inflation between 1962 and 1990, across OECD countries. There is a strong negative correlation between the two variables. Countries with little central bank independence in the mid-1980s were those where inflation was, on average, higher over the 30-year period.

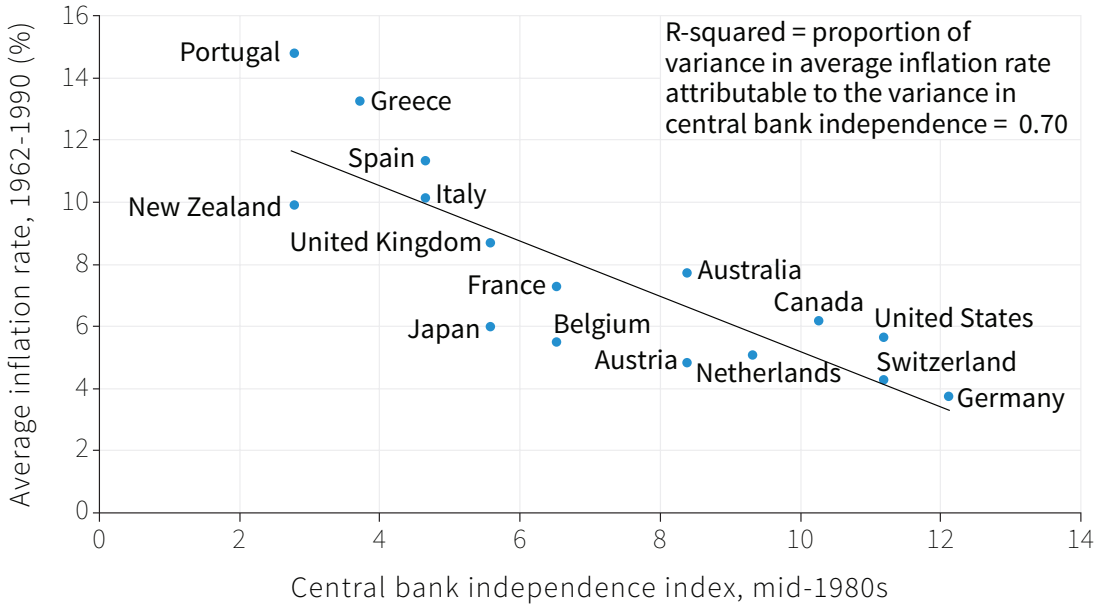


Figure 14.19 *Inflation and central bank Independence: OECD countries.*

Source: *CPI inflation: OECD. Independence of central bank: Grilli, Vittorio, Donato Masciandaro, Guido Tabellini, Edmond Malinvaud, and Marco Pagano. 1991. 'Political and Monetary Institutions and Public Financial Policies in the Industrial Countries.' Economic Policy 6 (13): 341–92.*

We can't conclude from this correlation how, or even if, central bank independence limited inflation, but many suspected that central bank independence would make it easier to control inflation. As a result, the high-inflation countries granted much more independence to their central bank, with a low inflation target embedded in official statutes.

New Zealand, which had high inflation in 1989, pioneered inflation targeting. Inflation fell and remained low. Other high-inflation countries soon followed, in particular Mediterranean countries like Portugal, Greece, Spain, Italy and France.

Evidently, those who believed that this correlation was a sign that central bank independence could control inflation were correct.

Under the policy of inflation targeting, whenever the economy was experiencing lower unemployment than the inflation-stabilising rate—moving to the north-east on a Phillips curve and on to a less favourable indifference curve—the central bank would raise the interest rate and dampen aggregate demand. Similarly, following a fall in aggregate demand (as a result of a fall in business confidence, for example) and facing the threat of recession, the central bank would cut the interest rate and bring the economy back toward its inflation target. We described the actions of the Federal Reserve in these terms in Figure 14.16.

Figure 14.20 shows the Phillips curve and indifference curves for an economy with an inflation-targeting central bank. The economy has stable inflation at point X, where inflation is at the policymaker's 2% target and unemployment is 6%. The inflation-stabilising rate of unemployment will be different in different countries. For example, during the 2000s, it was estimated at 5.9% in the UK, and 7.7% in Germany.

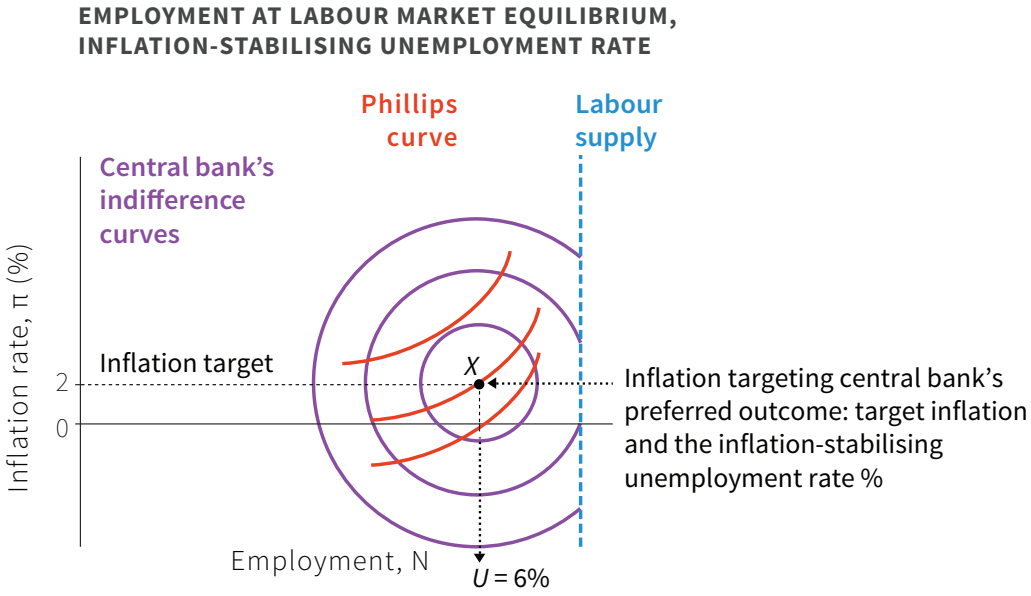


Figure 14.20 *The economy's stable inflation unemployment rate.*

If an aggregate demand shock reduces unemployment below 6%, inflation rises along the Phillips curve. In response, the central bank would raise the interest rate to reduce aggregate demand and raise unemployment. Likewise, if inflation should fall below target, the central bank will lower the interest rate to put upward pressure on inflation. Unless the central bank acts promptly, a wage-price spiral can begin, with the Phillips curve shifting upward.

The commitment of central banks to an inflation target helps explain why the third oil shock in the 2000s did not provoke a return to the high inflation of the 1970s. The commitment meant that even if the inflation rate rose temporarily, no one expected it to last because the central bank was committed to preventing it. Therefore there was no reason for an expected inflation spiral to begin.

14.12 ANOTHER REASON FOR RISING INFLATION AT LOW UNEMPLOYMENT

Why is there a trade-off in the economy between unemployment and inflation? So far the answer is that when unemployment is high in the economy, employees face a high cost of job loss, and employers will be able to get workers to work conscientiously at a lower wage than would be the case when unemployment is lower.

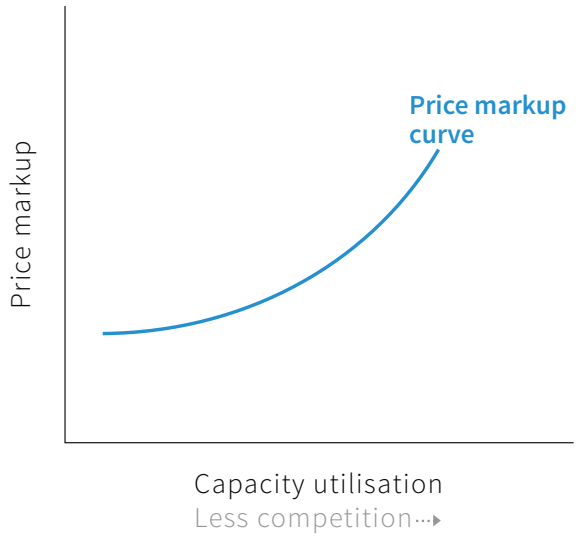


Figure 14.21 *Price responses to rising employment and capacity utilisation.*

But there is a second reason for the relationship between low unemployment and high inflation. In Figure 14.21, the horizontal axis shows the degree of capacity utilisation in the economy. When capacity utilisation rises as we move to the right along the horizontal axis, fewer machines are idle, there are fewer empty tables in restaurants, and other indicators (for example, more people working overtime shifts) show a reduction of spare capacity in factories and shops. In Unit 13 we explained the response of firms to rising capacity utilisation: they increase investment to expand their ability to meet orders.

However, building new plants and installing new equipment takes time. Meanwhile, at current prices, firms have more orders than they can fill. Economists say they are capacity-constrained. They lose nothing by raising prices in these conditions. Moreover their competitors—firms producing similar products—are capacity-constrained too, so the firms face less competition, meaning that their demand curves are now steeper (less price-elastic). So all firms will tend to respond to higher capacity utilisation by raising the markup of prices above costs, and this will kick off a wage-price spiral.

14.13 CONCLUSION

Voters want the economy to operate with low unemployment and low inflation. We will see in the next unit that some countries do this much better than others. But as we have seen here, achieving this outcome is not easy. And politicians will be tempted to make promises about delivering lower unemployment, for example, without taking account of the future inflation that would occur were they to be successful. This explains why monetary policy in the hands of independent central banks became the policy setup of choice in the 1990s.

The new macroeconomic policy framework of targeting inflation seemed to be working well when tested by the oil shock in the 2000s. Then came the global financial crisis, which rocked the consensus about the ability of the new policy framework to safeguard the economy against severe downturns. Like the Phillips curve itself, the success of the great moderation may not have been stable; it may have been living on borrowed time. We will explain why in Unit 17.

CONCEPTS INTRODUCED IN UNIT 14

Before you move on, review these definitions:

- *Inflation, Deflation, Disinflation, Expected inflation*
- Interest rates: *real, nominal, market, policy*
- Conflicting claims on output
- *Bargaining power* of firms over consumers, workers over firms
- *Phillips curve*, shifting Phillips curve
- *Bargaining gap*
- Policymaker's preferences
- *Monetary policy* transmission, exchange rate channel
- *Exchange rate*
- *Quantitative easing*
- *Supply shocks, Demand shocks*
- Central bank independence
- *Inflation target*
- *Capacity-constrained* firms

Key points in Unit 14

Voters dislike rapid price changes

Voters don't like inflation or deflation, and they don't like unemployment.

There is a short run trade-off between inflation and unemployment

This is known as the *Phillips curve*. The target combination of output and inflation on the Phillips curve that policymakers attempt to implement will be affected by whether voters' preferences are inflation-averse or unemployment-averse.

Monetary policy

This is used to adjust the level of aggregate demand in response to shocks. The central bank changes the policy interest rate in order to affect market interest rates and interest-sensitive spending (new housing, consumer durables, and machinery and equipment). When the economy is at the zero lower bound, the policy rate cannot be lowered and monetary policy relies on quantitative easing.

Shocks in the short-run

In the short run:

- Demand-side shocks, such as an investment slump, or greater demand for exports move the economy along the Phillips curve.
- Adverse supply-side shocks, such as an oil price increase, shift the Phillips curve upward, so that each level of unemployment is associated with a higher level of inflation.
- Favourable supply-side shocks, such as an improvement in technology, shift the Phillips curve downward.

Expected inflation

Workers and the owners of firms form expectations about inflation prior to wage and price setting.

Stable inflation

At the rate of unemployment at which claims for shares in output per capita are consistent, the rate of inflation is stable. If unemployment is below the stable inflation rate, then, as expected inflation is updated, the Phillips curve shifts up each period and inflation will be rising.

Inflation targeting

Since the 1990s, policymakers in many countries have targeted a specific inflation rate. These governments typically delegated responsibility for stabilisation of the economy around the inflation target to central banks. Sometimes central banks have become independent of the government.

14.14 EINSTEIN

The real interest rate

From Unit 11, the interest rate tells you how many dollars (or euros, pounds, or the currency you use) you will have to pay in the future in exchange for borrowing $1 today. If you are a lender, it tells you how many dollars you will receive in the future by giving up the use of $1 today.

The interest rates that you see quoted in bank windows or bank websites are nominal interest rates. That is to say that they do not take inflation into account. If you are a lender, what you really want to know is how many goods you will get in the future in exchange for the goods you don't consume now. If you are a borrower, what matters is how many goods you will have to give up in the future to pay the interest, rather than the total interest measured in dollars. The opportunity cost of the loan is the goods you have to give up, not the money you have to give up. To make this distinction, you need to take account of inflation.

Households and firms make decisions based on real interest rates. Firms will judge which investment projects are worth undertaking using real interest rates, and lenders will charge a higher level of interest on their loans if inflation is expected to erode their lending margins in the future.

The equation for the real interest rate is known as the *Fisher equation*, after Irving Fisher, whose physical model of the economy we saw in Unit 2. The Fisher equation states that *the real interest rate (% per annum) equals the nominal interest rate (% per annum) minus the inflation expected over the year ahead*:

$$r = i - \pi^e$$

When evaluating an investment project, the expected inflation rate needs to be taken into account. We can see that when prices are expected to fall over the year ahead—that is, expected inflation is negative—it raises the real interest rate. At the higher real interest rate, some investment projects that would have been undertaken in the absence of forecast deflation are ruled out.

The profit curve with imported materials

In the Einstein in Unit 9, we explained the steps to show how the profit curve for the economy as a whole results from the decisions of individual firms. Here we take a short cut and go straight to the economy as a whole. An individual firm in the economy uses as inputs both the products of other firms in the economy and

imported products. The cost of the inputs will be affected by wage costs and imported materials costs. Once we aggregate all the firms, we have only two types of cost: labour and imported materials.

The firm's costs in addition to wages are imported materials. The firm's marginal (and average) cost is its unit labour cost (ulc) plus its unit imported materials cost (umc). So unit costs are:

$$uc = umc + ulc$$

Output per worker is q, so:

$$ulc \equiv \frac{W}{q}$$

This is wage per worker divided by output per worker. We define the markup, μ, as the share of the price that represents profits to the firm (what is left over after subtracting unit costs):

$$\mu \equiv \frac{(P - umc - ulc)}{P}$$
$$\equiv \frac{1 - umc}{P} - \frac{ulc}{P}$$

Suppose the price per unit is \$5, imported materials cost \$1 per unit and labour costs \$2.50 per unit. Then the markup is:

$$\mu = 1 - \frac{1}{5} - \frac{2.5}{5} = 0.3$$

which is 30%. Substituting $ulc = W/q$ gives us:

$$\mu \equiv 1 - \frac{umc}{P} - \frac{(W/q)}{P}$$

Multiplying each side by q and rearranging, we get the *profit curve*:

$$\frac{W}{P} \equiv q\left(1 - \mu - \frac{umc}{P}\right)$$

This shows that the wage per worker is equal to output per worker q minus profits per worker that go to the owner, μ, minus imported materials costs that go to foreign producers umc/P. Any increase in unit materials costs such as a rise in the price of oil will shift the profit curve down.

An equivalent but alternative version of the markup equation is provided in the next section.

The markup price-setting equation for the firm

When explaining the process of inflation, it is useful to have an equation describing explicitly how firms set and change their prices. As we saw in the Einstein in Unit 9, the price set by firms is a markup on its costs, where the markup depends on the extent of competition in the product market.

We defined the markup as the share of the price that was profit to the firm. We can also define the markup m as how much the firm charges above the cost per unit. For m, the *markup price-setting equation* looks like this:

$$\begin{aligned} P &= (1+m)(\textit{unit costs}) \\ &= (1+m)(umc+ulc) \end{aligned}$$

Where the markup is m, umc is the unit cost of materials and ulc is the unit cost of labour.

The markup price-setting equation says that if unit costs are \$3.00 and the markup m is 10%, the price will be \$3.30. So the extra \$0.30 charged above costs is equal to 10% of those costs. If we want to know in this case, we ask what the extra \$0.30 is as a share of the total price, rather than as a share of the cost. Then $\mu = \$0.30/\$3.30 = 0.09$ or 9%.

One advantage of using m is that it makes it easy to see that if the markup is fixed then a rise in unit costs of, for example, 5% must imply a price rise of 5%. This follows directly from the markup price-setting equation above.

We can also ask what happens to P when just one part of the costs rise, such as the imported materials cost. Assuming m remains constant, the percentage change in the price is is equal to the percentage change in total unit costs:

$$\begin{aligned} \frac{\Delta P}{P} &= \frac{(1+m)\Delta(umc+ulc)}{(1+m)(umc+ulc)} \\ &= \frac{\Delta umc}{(umc+ulc)} + \frac{\Delta ulc}{(umc+ulc)} \end{aligned}$$

We now divide both the numerator and the denominator of the first term on the right hand side by umc, and the second term by ulc:

$$\frac{\Delta P}{P} \equiv \frac{(\Delta umc/umc)}{(umc+ulc)/umc} + \frac{(\Delta ulc/ulc)}{(umc+ulc)/ulc}$$

This is equivalent to:

$$\frac{\Delta P}{P} \equiv \frac{\Delta umc}{umc} \times \frac{umc}{(umc+ulc)} + \frac{\Delta ulc}{umc} \times \frac{ulc}{(umc+ulc)}$$

In words, the percentage change in P is equal to the percentage change in umc times umc's share of unit costs, plus the percentage change in ulc times ulc's share of unit costs. For example, suppose the markup is 60% and unit cost is \$5, of which \$4 is

labour cost and $1 is imported materials, so the price is $P = 1.6 \times \$5 = \$8$. Wages are 80% of the cost, so if wages go up 10% then the price will rise by $80\% \times 10\% = 0.8\%$. In this example, unit costs rises to $\$4.4 + \$1 = \$5.4$ and the price rises to $P = 1.6 \times \$5.4 = \$8.64$ (a rise of 8%). Equally, if the price of imports, such as oil, were to rise by 10% then the price would rise by $20\% \times 10\% = 2\%$.

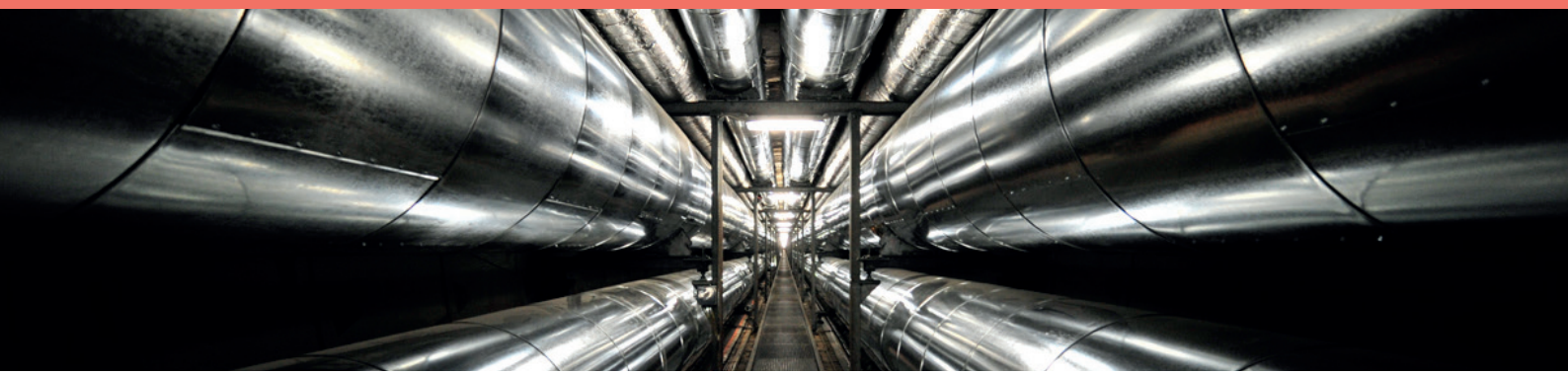
14.15 READ MORE

Bibliography

1. Bank of England. 2015. 'Three Centuries of Macroeconomic Data.'
2. Bordo, Michael, Barry Eichengreen, Daniela Klingebiel, and Maria Soledad Martinez-Peria. 2001. 'Is the Crisis Problem Growing More Severe?' *Economic Policy* 16 (32): 52–82.
3. Federal Reserve Bank of St. Louis. 2015. 'FRED.'
4. Friedman, Milton. 1968. 'The Role of Monetary Policy.' *The American Economic Review* 58 (1): 1–17.
5. Grilli, Vittorio, Donato Masciandaro, Guido Tabellini, Edmond Malinvaud, and Marco Pagano. 1991. 'Political and Monetary Institutions and Public Financial Policies in the Industrial Countries.' *Economic Policy* 6 (13): 341–92.
6. Jahan, Sarwat. 2012. 'Inflation Targeting: Holding the Line.' *International Monetary Fund Finance & Development*.
7. OECD. 2015. 'OECD Statistics.'
8. Phillips, A W. 1958. 'The Relation between Unemployment and the Rate of Change of Money Wage Rates in the United Kingdom, 1861-1957.' *Economica* 25 (100): 283.
9. *The Economist*. 2003. 'Bush's Push.' January 6.
10. *The Economist*. 2013. 'In Dollars They Trust.' April 27.
11. *The Economist*. 2013. 'Controlling Interest.' September 21.
12. US National Archives. 2012. '1789-2012 Presidential Elections.' US Electoral College.
13. Walton, David. 2006. 'Has Oil Lost the Capacity to Shock?' *Oxonomics* 1 (1): 9–12.

Beta January 2016 version

TECHNOLOGICAL PROGRESS, UNEMPLOYMENT AND LIVING STANDARDS IN THE LONG RUN



HOW LONG-TERM TRENDS IN DIFFERENCES IN LIVING STANDARDS AND UNEMPLOYMENT BETWEEN COUNTRIES ARE THE RESULT OF TECHNOLOGICAL PROGRESS, INSTITUTIONS AND POLICIES

- The increasing use of machinery and other capital goods in production, along with technological progress made possible by increasing knowledge, have been the foundation for increased living standards in the long run
- The "creative destruction" of older ways of producing goods and organising production has meant continuous job loss as well as job creation, but not higher unemployment in the long run
- A country's economic institutions and policies may be evaluated by their capacity to keep involuntary unemployment low and to sustain increases in real wages
- Many successful economies have provided extensive forms of co-insurance against the job losses arising from both creative destruction and competition from other economies, so that most citizens of these nations welcome both technological change and the global exchange of goods and services
- The main difference between high-performing economies and laggards is that the institutions and policies of high performers work so that the main actors are incentivised to increase the size of the pie, rather than fighting over the size of their slice

Institute for New Economic Thinking
UCL SciencesPo Azim Premji University
FRIENDS PROVIDENT Foundation

In 1412 the city council of Cologne prohibited the production of a spinning wheel by a local craftsman because it feared unemployment among textile manufacturers that used the hand spindle. In the 16th century, new ribbon-weaving machines were banned in large parts of Europe. In 1811, in the early stage of the industrial revolution in England, the Luddite movement protested forcefully against new labour-saving machinery, such as spinning machines that allowed one worker to produce the amount of yarn previously produced by 200 workers. The movement was led by a young unskilled artisan, Ned Ludd, who allegedly destroyed the spinning machines.

The Swiss economist Jean-Charles-Léonard de Sismondi (1773-1842) contemplated a new world "where the King sits alone on his island, endlessly turning cranks to produce, with automatons, all that England now manufactures". The increasing use of information technology has caused contemporary economists, including Jeremy Rifkin, to voice the same fears.

Sismondi and Rifkin made plausible arguments. But, as we saw in Unit 1, as a result of labour-saving innovations many countries moved to the upward part of the hockey stick and experienced sustained growth in living standards. Workers were paid more—remember the real wage hockey stick from Unit 1, and the data on wages in Unit 3—and the "end of work" hasn't happened yet. (Although in 1935 Bertrand Russell, a philosopher, expressed anticipation rather than fear of the end of work, arguing that: "[T]here is far too much work done in the world, that immense harm is caused by the belief that work is virtuous, and that what needs to be preached in modern industrial countries is quite different from what always has been preached.")

Technological progress has not created rising unemployment rates. Instead it has raised the lowest wage that firms can pay while still covering their costs. As a result, technological progress expands the resources the firm has to invest in increasing production. By focusing only on the destruction of jobs, those who worry about the end of work have ignored the fact that labour-saving technological progress also helps to create them.

In most economies for which data is available, at least 10% of jobs are destroyed every year, and about the same number of new ones are created. Every day, in France or the UK, a job is destroyed and another one created every 14 seconds. This is part of the creative destruction process at the heart of capitalist economies that we described in Unit 1 and Unit 2.

Those who lose their jobs bear substantial costs in the *short run*. The short run may not seem very short to them: it can last years or even decades. Those who benefit may be the children of the handloom weaver displaced by the power loom; the children of the unemployed typist who was displaced by the computer. They benefit by finding a job in an occupation that is more productive than the job their parents did, and they may share in the benefit from the new goods and services that are available because the power loom or the computer exist. The destructive part of creative destruction affects occupations that may often be concentrated in particular regions, with large losses of wages and jobs. Families and communities who are the losers

often take generations to recover. Like "short run", the term "average" often hides the costs to workers displaced, and communities destroyed, by the introduction of new technologies.

Today, for example, information and communication technology (ICT) is reshaping our societies. ICT is replacing much routine labour, in many cases further impoverishing the already poor. People who would have previously anticipated rising living standards have fewer job opportunities.

Nevertheless, most people benefit from the fall in prices due to the new technology. For better or worse, creative destruction as a result of technological progress is part of the dynamism of the capitalist economic system. And while lives have been disrupted and the environment increasingly threatened by this dynamism, the introduction of improved technologies is also the key to rising living standards in the long run. We shall see that:

- Technological change is constantly putting people out of work
- But the countries that have avoided high levels of unemployment are among those in which the productivity of labour has increased the most

Figure 15.1 shows unemployment rates for 16 OECD countries from 1960 to 2012:

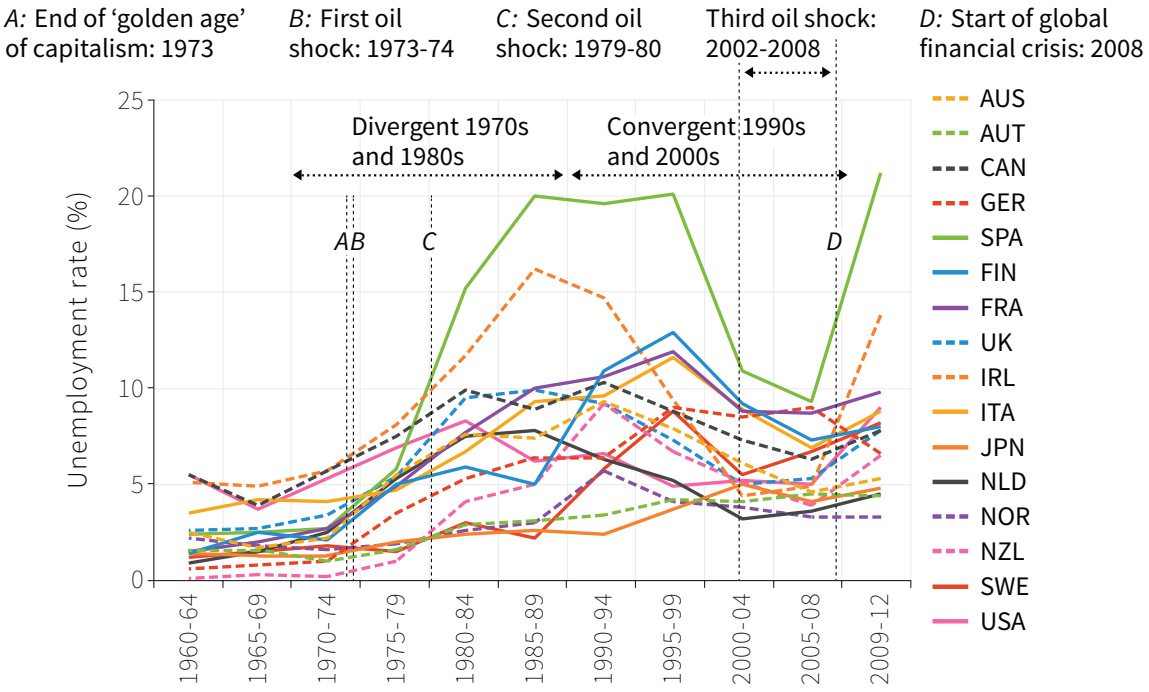


Figure 15.1 *Unemployment rates in selected OECD countries (1960-2012).*

Source: *Data from 1960-2004: Howell, David R, Dean Baker, Andrew Glyn, and John Schmitt. 2007. 'Are Protective Labor Market Institutions at the Root of Unemployment? A Critical Review of the Evidence.' Capitalism and Society 2 (1). Data from 2005 to 2012: OECD. 2015. Harmonised unemployment rates, 'OECD Statistics.'*

Unemployment rates were low and quite similar in the 1960s, and then diverged in the 1970s, reflecting in part the different country responses to the oil shocks described in Unit 14. Of these countries, only Japan (JPN), Austria (AUT) and Norway (NOR) have unemployment rates that stayed below 6% over the entire period. In Spain (SPA) unemployment was around 20% from the mid 1980s to the end of the 1990s. It then halved in the 2000s before jumping back above 20% following the financial crisis and eurozone crisis from 2009. In this respect Germany (GER) is unusual: unemployment fell in the years following the global financial crisis.

While there has been no upward trend in unemployment rates over the long run, there have been two important developments in the labour market that have accompanied the growth in living standards. As we saw in Unit 3, average annual hours worked by people with jobs have fallen and a larger fraction of adults are working for pay. This has been mainly due to the rise in the proportion of women who do paid work.

The patterns of unemployment in Figure 15.1 are not explained by national differences in the rate of innovation, or waves of innovation over time. They reflect differences in the institutions and policies in force in the countries at different times.

How have living standards improved over the long run, as production has become more capital intensive, without producing mass unemployment? We begin by studying the accumulation of capital (the increasing stock of machinery and equipment), as well as infrastructure (such as roads and ports), which has always been fundamental to the dynamism of capitalism.

DISCUSS 15.1: WEALTH AND LIFE SATISFACTION

As we saw in Unit 3, technological progress increases your productivity per hour. This means that by working the same number of hours you could thus produce and consume more; or you can produce and consume the same amount of goods while working fewer hours and enjoying more free time.

The economist Oliver Blanchard argues that the difference in output per capita between the US and France is partially due to the fact that relative to those in the US, the French have used some of the increase in productivity to enjoy more free time rather than raise consumption (see Figures 3.1 and 3.2).

1. Would you expect life satisfaction to be lower or higher in a country that has lower GDP per capita due to lower hours worked, as is the case in France relative to the US? Explain your answer.
2. Considering your answer above, which country—France or the US—would you prefer to live in and why?

15.1 TECHNOLOGICAL PROGRESS AND LIVING STANDARDS

In Unit 2 we saw how firms could earn Schumpeterian *innovation rents* by introducing new technology. Firms that fail to innovate (or copy other innovators) are unable to sell their product for a price above the cost of production, and eventually fail. This process of *creative destruction* led to sustained increases in living standards on average because *technological progress* and the accumulation of *capital goods* are complements: each provides the conditions necessary for the other to proceed.

- *New technologies require new machines:* The accumulation of capital goods is a necessary condition for the advance of technology, as we saw in the case of the spinning jenny.
- *Technological advance is required to sustain the process of capital goods accumulation:* It means that the introduction of increasingly capital-intensive methods of production continue to be profitable.

The second point above needs explanation. Start with the production function that we used in Units 2 and 3. We discovered that output depends on labour input—and that the function describing this relationship shifts upward with technological progress, so that the same amount of labour now produces more output. In Unit 3 the farmer had a fixed amount of land: we assumed the amount of capital goods was fixed. But, as we have seen, the amount of capital goods which the modern worker uses is vastly greater than that used by farmers in the past.

CREATIVE DESTRUCTION

Joseph Schumpeter's name for the process by which old technologies (and the firms that do not adapt) are swept away by the new, because they cannot compete in the market.

- The failure of unprofitable firms is creative, in his view, because it releases labour and capital goods for use in new combinations.

Now include capital goods (machinery, equipment and structures) explicitly. If you look at the horizontal axis in Figure 15.2 you will see that it records the amount of capital goods per worker. This is a measure of what is called the capital intensity of production. On the vertical axis, we have the amount of output per worker, or what is called *labour productivity*.

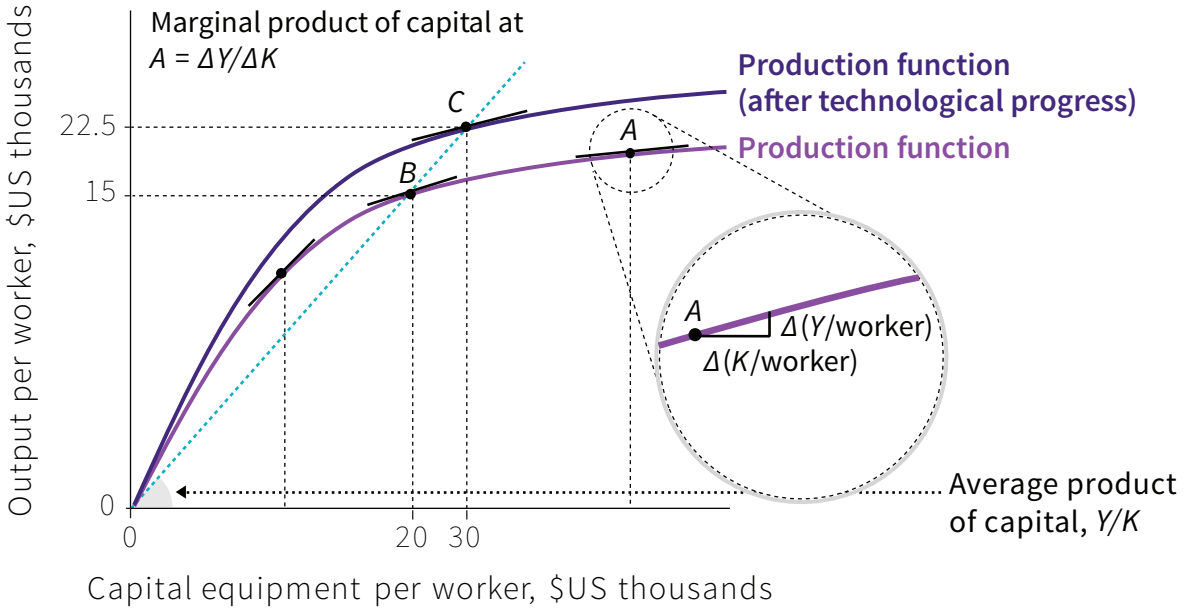


Figure 15.2 *The economy's production function and technological progress.*

As was the case in Unit 3 the production function describes diminishing returns: as the worker works with more capital goods output increases, but at a diminishing rate (Charlie Chaplin showed in the 1936 film Modern Times that there is a limit to the number of machines a worker can make use of). This means that, with increasing quantities of capital goods, we have a diminishing marginal product of capital goods. The slope of the production function at each level of capital per worker shows the marginal product of capital: it shows how much output increases if capital equipment per worker increases by one unit. The magnified section at point A in Figure 15.2 shows how the marginal product of capital is calculated: note that *Y/worker* is used as shorthand for output per worker, and the marginal product of capital (MPK) is $\Delta Y/\Delta K$. The marginal product of capital at each level of capital per worker is the slope of a tangent to the production function at that point. Previous Leibniz supplements showed how to use calculus to calculate this for a given production function. Take a moment to have another look at the relevant Leibniz supplements: this one from Unit 2 and this one and this one from Unit 3.

We can see from Figure 15.2 that the marginal product of capital is falling as we move along the production function. The shape of a production function that exhibits diminishing returns to capital is called *concave*: concavity captures the fact that output per worker increases with capital per worker, but less than proportionally.

We plot data from the year 1990 on capital per worker and output per worker in Figure 15.3, for 57 countries. The dotted grey line sketches a concave production function, with diminishing marginal productivity of capital, like the one in Figure 15.2.

The data for all the countries is shown in US dollars at 1985 prices.

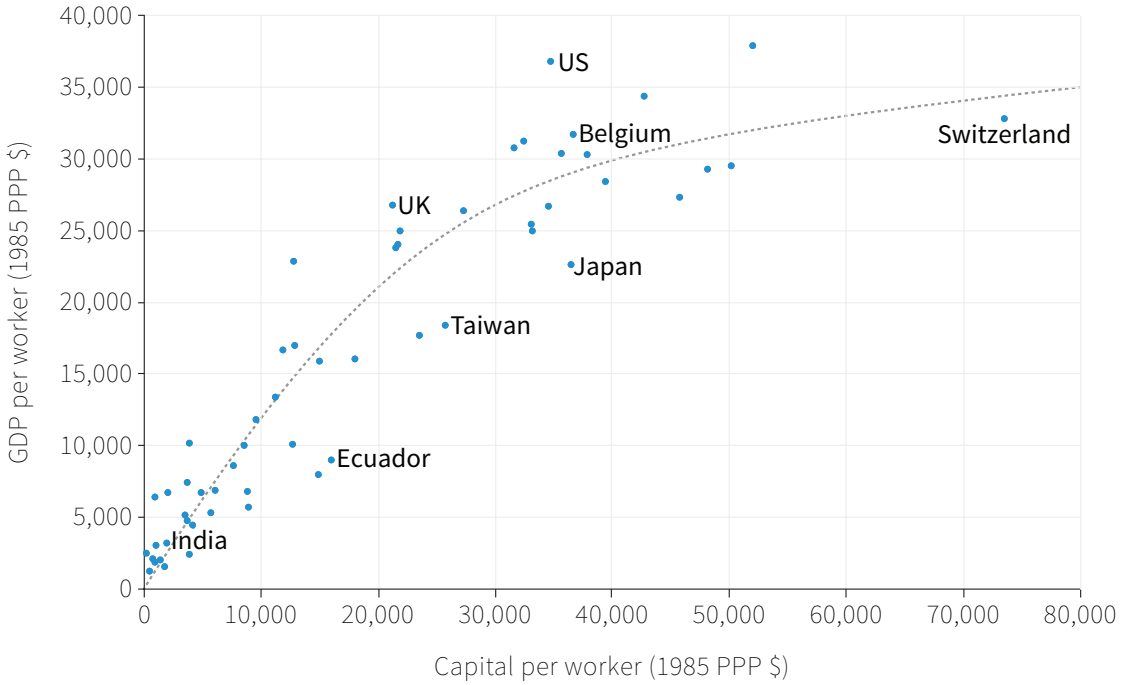


Figure 15.3 *Capital and GDP per worker across countries in 1990.*

Source: *Allen, Robert C. 2012. 'Technology and the Great Divergence: Global Economic Development since 1820.' Explorations in Economic History 49 (1): 1–16.*

If accumulation of capital proceeded in the absence of technological progress, an economy might progress in stages from being like Ecuador, becoming like Belgium, and eventually even becoming Switzerland. As this hypothetical economy progresses along the dotted line of the production function, the contribution of additional capital goods to increased production would eventually become so small that capital accumulation could no longer drive growth in labour productivity and living standards. Notice that when the economy gets to the position of Switzerland, the production function is relatively flat and the marginal productivity of capital (increase in GDP per worker for each unit increase in capital goods per worker) is low. The process of growth would slow and could eventually cease, an outcome called a *stationary state*. In the 19th century, in his book *The Principles of Political Economy*, John Stuart Mill, a philosopher and economist, welcomed this prospect as "a very considerable improvement on our present condition".

DISCUSS 15.2: A STATIONARY STATE

1. Considering the empirical relationship in Fig 15.3 above, do you agree with John Stuart Mill that being at or close to the position of Switzerland on this curve is a desirable outcome?
2. Discuss the advantages and disadvantages of a stationary state.

We have not reached the stationary state, because technological progress rotates the production function upward as we saw in Unit 3, and as shown in Figure 15.2.

With the same amount of capital goods per worker, the new technology allows the production of a larger quantity of goods with a given amount of labour. This is called *technological progress*, and it increases the average productivity of labour.

Technological progress also increases the marginal product of capital (for any given level of capital goods per worker), offsetting the diminishing marginal productivity that is characteristic of a concave production function. To see this in Figure 15.2, we mark point *B* on the original production function, where capital per worker is \$20,000 and output per worker is \$15,000. To put this in context, these figures roughly correspond to the Japanese economy at the end of the 1970s. On the production function, after technological progress we mark point *C*, at which capital per worker has risen to \$30,000 and output per worker has risen to \$22,500. We can also see that, in this case, the slope of the production function at points *B* and *C* is the same. Hence, even though capital per worker is higher at point *C*, the marginal productivity of capital (as shown by the slope of the production function) is the same as at point B. In other words, for a given level of capital per worker, technological progress has improved the marginal productivity of capital. The new technology makes it profitable to increase capital intensity and move to higher output per worker at point *C*.

New technology can also refer to new ways of organising work: remember that a technology is a set of instructions for combining inputs to make output. The managerial revolution in the early 20th century called *Taylorism* is a good example: labour and capital equipment were reorganised in a streamlined way, and new systems of supervision were introduced to make workers work harder. More recently, the information technology revolution allows one engineer to be connected with thousands of other engineers and machines all over the world. The ICT revolution therefore rotates the production function upward, increasing its slope at every level of capital per worker.

In Figure 15.2, you can see a dotted blue line from the origin through the production functions for the old and new technologies. The slope of this line tells us the amount of output per unit of capital goods at the point where it hits the production function: it is the amount of output per worker divided by the capital goods per worker. From the diagram, we note that points *B* and *C* on the two production functions have the same output per unit of capital goods.

To see how technological progress and capital accumulation shaped the world, we focus on the countries that have been the technology leaders. Britain was the technological leader from the Industrial Revolution until the eve of the first world war, when the US took over leadership. Figure 15.4 has the same axes as Figure 15.3, and once again shows the world's production function in 1990 (the dashed line). We can now look at the path traced out over time by the UK and the US. Looking first at Britain, the data begins in 1760 (the bottom corner of the chart) and ends in 1990

with much higher capital intensity and productivity. We can show the same points in the familiar hockey stick chart for GDP per worker, in the bottom right-hand side of the diagram. As Britain moved up the hockey stick over time, both capital intensity and productivity rose. In the US, productivity overtook the UK by 1910 and has remained higher since. In 1990 the US had higher productivity and capital intensity than the UK.

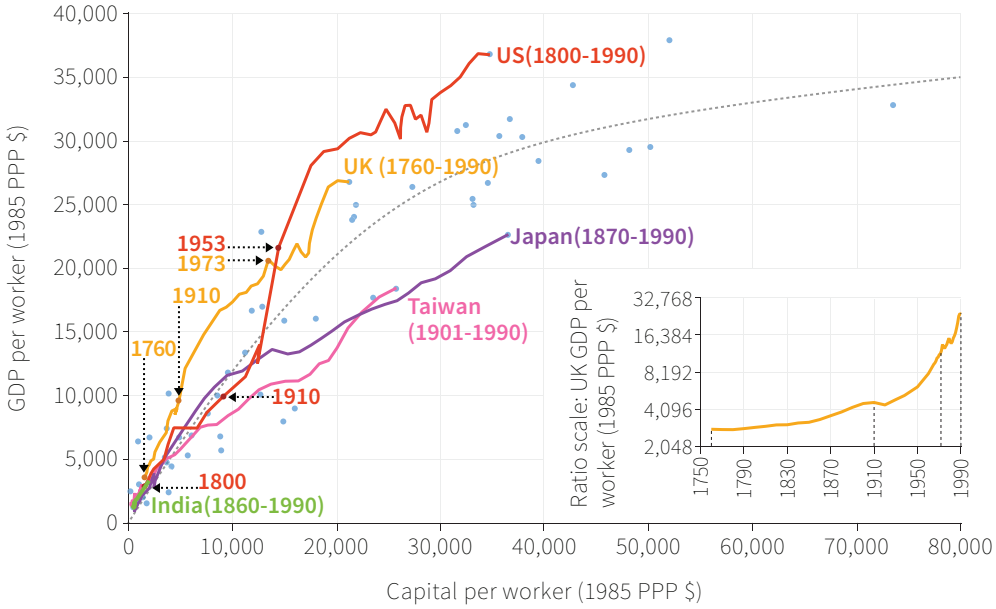


Figure 15.4 *Long-run growth trajectories of selected economies.*

Source: *Allen, Robert C. 2012. 'Technology and the Great Divergence: Global Economic Development since 1820.' Explorations in Economic History 49 (1): 1–16.*

Figure 15.4 shows that the countries that are rich today have seen labour productivity rise over time as they became more capital-intensive. For example, if we look at the US, capital per worker rose from $4,325 in 1880 to $14,407 in 1953, and $34,705 in 1990 (all measured in 1985 US dollars). (John Habakkuk, an economic historian, has argued that wages were high for factory workers in the US in the late 19th century because they had the option to move west: therefore the factory owners had the incentive to develop labour-saving technology.) Alongside this increase in capital intensity, US labour productivity rose from $7,400 in 1880 to $21,610 in 1953, to $36,771 in 1990 (all measured in 1985 US dollars).

Productivity growth has reduced labour input per unit of output: the fear of the Luddites and the forecasts of the "end of work" authors was that this would cause permanent job loss.

From Figure 15.4 it is clear that the historical paths traced out by Britain and the US, which reflect the combination of capital accumulation and technological progress, are not as curved as the production function in Figure 15.2, or the labour productivity

across countries in a single year (as shown by the dotted line in Figure 15.4). The countries that were the leaders in the introduction of new technology moved along a path similar to the blue dotted line between *B* and *C* in Figure 15.2.

We know from Unit 1 that other countries moved up the hockey stick at very different times. Three other countries are shown in Figure 15.4: Japan, Taiwan and India. The paths of these three countries show that moving along the hockey stick curve of living standards requires capital accumulation and the adoption of new technology. Their paths are also not as curved as (that is, less concave than) the 1990 production function. Notice that, by 1990, capital per worker in Japan was not only higher than in the US, but almost twice as high as in Britain. Japan had reached this level in less than half the time it took Britain. Taiwan in 1990 was also more capital-intensive than Britain. The lead in mass production and science-based industries that the US had established was eroded as other countries invested in education and research, and adopted American management techniques, as this paper demonstrates.

Interpreting Figures 15.3 and 15.4 using the model of the production function in Figure 15.2 shows that *countries adopted more capital-intensive methods of production as they became richer, and labour became better paid*.

An early stage of this process was the introduction of the labour-saving, energy-using spinning jenny studied in Unit 2. To summarise:

- *Technological progress shifted the production function up*.
- *This offset the diminishing marginal returns to capital:* Capital productivity, measured by the slope of a ray from the origin, remained roughly constant over time in the technology leaders.

Technological progress played a crucial role in preventing diminishing returns from bringing to an end the long-run improvement in living standards resulting from the accumulation of capital goods.

15.2 TURBULENCE AND THE JOB CREATION AND DESTRUCTION PROCESS

Labour-saving technological progress of the type illustrated in Figures 15.2 and 15.4 allows more to be produced with a given amount of labour, and it also contributes to the expansion of production. It compensates for some of the jobs it has destroyed, and may even create more jobs than previously existed. When more jobs are created than destroyed in a given year, employment increases. When more jobs are destroyed than created, employment decreases.

We know that at any moment there are some people who are involuntarily unemployed. They would prefer to be working, but don't have a job. The number of unemployed is a stock, measured without a time dimension.

Their numbers change from day to day, or year to year, as some of the jobless are hired (or give up seeking work), other people lose a job, and yet others decide to seek work for the first time (young people leaving school or university, for example). Those without work are sometimes called the "pool" of the unemployed: people getting a job or ceasing to look for one flow out of the pool, while those who lose their jobs flow into the pool. The number getting and losing jobs are a flow.

The total job reallocation process is the sum of job creation and destruction. Compared to that, the net growth of employment is typically small and positive.

For the countries shown in Figure 15.5, job destruction, job creation and net employment growth are shown—note that in the UK from 1980 to 1998, more jobs were destroyed than created: net employment growth was negative. Across a set of countries at different stages of development, and with different openness to international trade, we see a fairly similar rate of job reallocation. In most countries, about one-fifth of jobs are created or destroyed each year, in spite of widely varying rates of net employment growth.

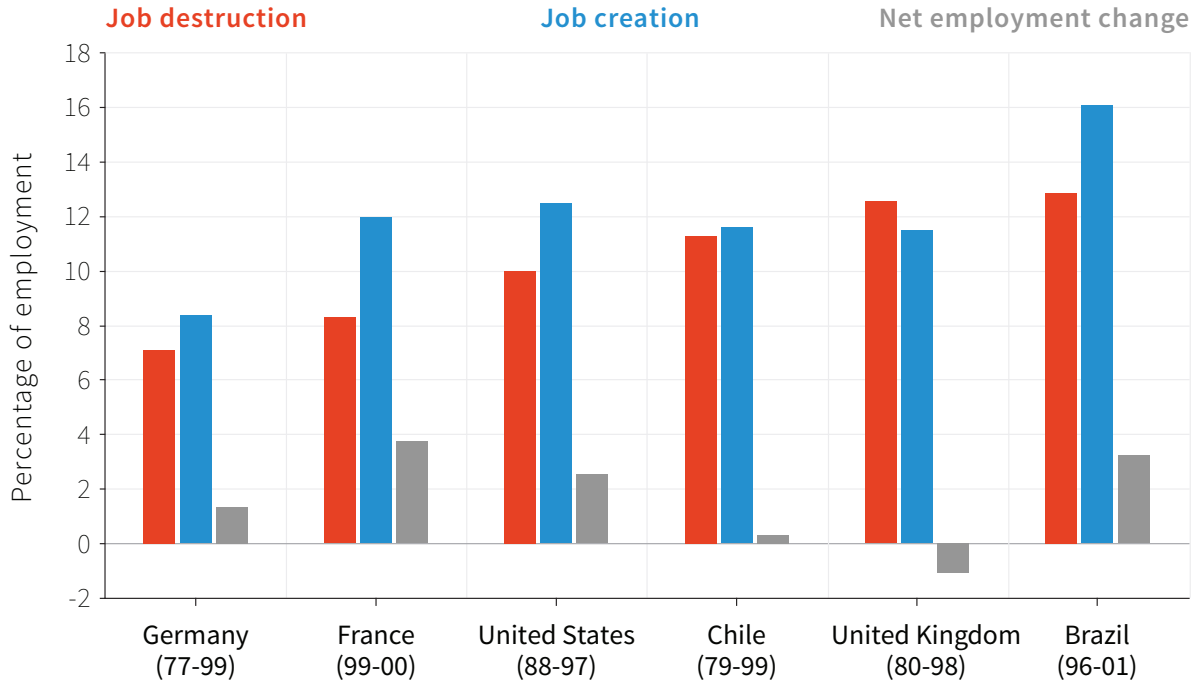


Figure 15.5 *Job destruction, job creation and net employment across countries.*

Source: *Haltiwanger, John, Stefano Scarpetta, and Helena Schweiger. 2014. 'Cross Country Differences in Job Reallocation: The Role of Industry, Firm Size and Regulations.' Labour Economics 26 (January): 11–25.*

Now think about an economic system in which 2% new jobs are created each year, and job destruction is banned (that is, the job destruction rate is zero). This economy would also see a net growth of employment of 2%. This is what a planner might seek to do. Figure 15.5 shows this is not the way a capitalist economy works in practice. There is no planner. Competition, and the prospect of gaining economic rents, mean that creating jobs often implies destroying other ones.

DISCUSS 15.3: SCHUMPETER REVISITED

1. In Unit 2 we introduced the way Joseph Schumpeter characterised capitalist economies by the process of "creative destruction". In your own words, explain what this term means.
2. Based on this definition and for an economy of your choice, give examples of destruction and creation and identify the winners and the losers.

To understand how job creation and destruction take place in an industry, we look at the impact of the information technology revolution in the retail sector since the 1990s in the US. The adoption of systems that electronically link cash registers to scanners, credit card processing machines, and to management systems for both inventories and customer relationships allowed tremendous increases in output per worker. Think of the volume of retail transactions handled per cashier working in a new retail outlet. Research shows that labour productivity growth in the retail sector was entirely accounted for by more productive new establishments (such as retail units or plants) displacing much less productive existing establishments (including older establishments of the same firm, as well as shops and plants owned by others, where jobs were lost).

We showed the massive expansion of employment in the US firm Walmart in Figure 15.2 of Unit 6. Walmart's growth was partly based on opening more efficient out-of-town stores, made possible by new retail and wholesale technologies.

For the manufacturing sector, detailed evidence collected from all the firms in the economy shows how productivity growth takes place through the creation and destruction of jobs inside firms, and by their entry and exit. The data for Finland in the years from 1989 to 1994, for example, shows that 58% of productivity growth took place within firms (similar to the Walmart example). The exit of low-productivity firms contributed to a quarter of the increase, and 17% more was contributed by the reallocation of jobs and output from low- to high-productivity firms.

The French construction industry provides another example of the reallocation of work from weaker to stronger firms. According to the French national institute of statistics, it was in firms with very low productivity (in the bottom 25%) that more of the jobs in the economy were destroyed than created. In these firms job creation

added 7.1% new jobs while job destruction took away 16.1% of jobs between 1994 and 1997, implying that employment in those firms shrank by 9.0%. In contrast, job creation exceeded destruction (17.1% against 11.8%) in the 25% of construction firms with the highest productivity.

15.3 JOB FLOWS, WORKER FLOWS AND THE BEVERIDGE CURVE

Jobs are created and destroyed by business owners and managers seeking to gain Schumpeterian innovation rents, and to respond to the pressure of competition in markets for goods and services. For most workers this means that nothing is permanent: in the course of a lifetime people move in and out of many jobs (often not by choice). Sometimes these are job-to-job moves, but sometimes they move in and out of unemployment too.

In Unit 5 we looked at the decisions by an employer (Bruno) and an employee (Angela) about her work hours and rent. Once Bruno's gun was replaced by a legal system and contracts, we saw that taking a job was a voluntary arrangement entered into for mutual gain. The balance of *bargaining power* may have been unequally distributed but the exchange was, nevertheless, voluntary.

When a worker leaves a job, it may be voluntary, but it can also be an involuntary temporary lay-off (dictated by product demand conditions facing the firm), or a redundancy (the job has been eliminated).

Jobs are also created, as can be seen by the movement of job destruction and creation in the US in Figure 15.6. Job creation is strongly *procyclical*: this means that it rises in booms, and falls during recessions. Conversely, job destruction is *countercyclical*: it rises during recessions (if the change in a variable was not correlated with the business cycle, it would be called *acyclical*). The next section will show how aggregate policies interact with those movements in job flows and worker flows.

This intense job reallocation process, and the ability of the government to provide co-insurance, led the English economist and politician Lord William Beveridge (1879-1963) to become the founding father of the UK social security system. He is also remembered among economists because, like Bill Phillips, they bestowed on Beveridge one of their highest honours: they named the *Beveridge curve* for him.

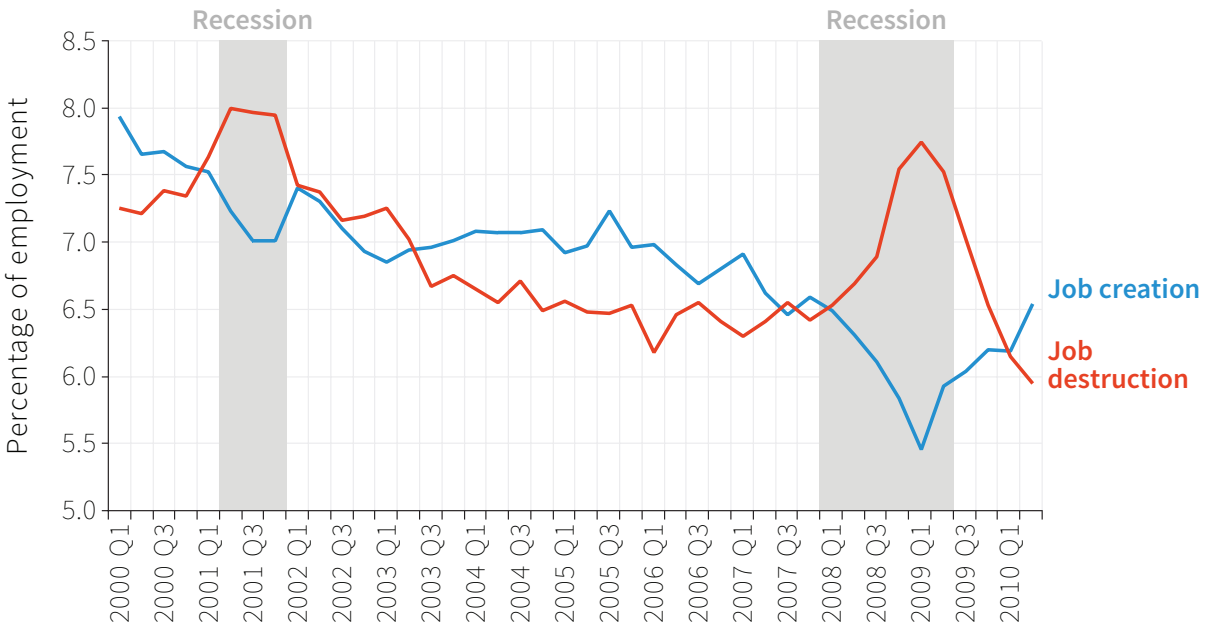


Figure 15.6 *Job creation and destruction during cycles in the United States (Q1 2000—Q2 2010).*

Source: *Davis, Steven J, R. Jason Faberman, and John C Haltiwanger. 2012. 'Recruiting Intensity during and after the Great Recession: National and Industry Evidence.' American Economic Review 102 (3): 584–88.*

The Beveridge curve

Beveridge had suggested a simple relationship between job vacancy rates (the number of jobs looking for workers) and the level of unemployment (the number of workers looking for jobs), expressed as a fraction of the labour force.

THE BEVERIDGE CURVE

The relationship between the unemployment rate and the job vacancy rate (each expressed as a fraction of the labour force):

- When unemployment is high, the vacancy rate is low
- When unemployment is low, the vacancy rate is high

The inverse relationship is named after William (Lord) Beveridge, the economist who discovered it.

Beveridge noticed that when unemployment was high, the vacancy rate was low; and when unemployment was low, the vacancy rate was high:

- *During recessions there will be high unemployment*: When the demand for a firm's product is declining or growing slowly, firms can manage with their current staff even if a few of them quit or retire. As a result they advertise few positions. In the same conditions of weak demand for firms' products, people will be laid off or their jobs entirely eliminated.

- *During booms the numbers of unemployed decline:* The number of vacant jobs posted by firms increases.

The downward-sloping relationship between the vacancy rate and the unemployment rate over the business cycle is illustrated in Figure 15.7.

Figure 15.7 shows two examples of what came to be called the Beveridge curve, for Germany and the US. Each dot represents a quarter, from Q1 2001 until Q2 2015.

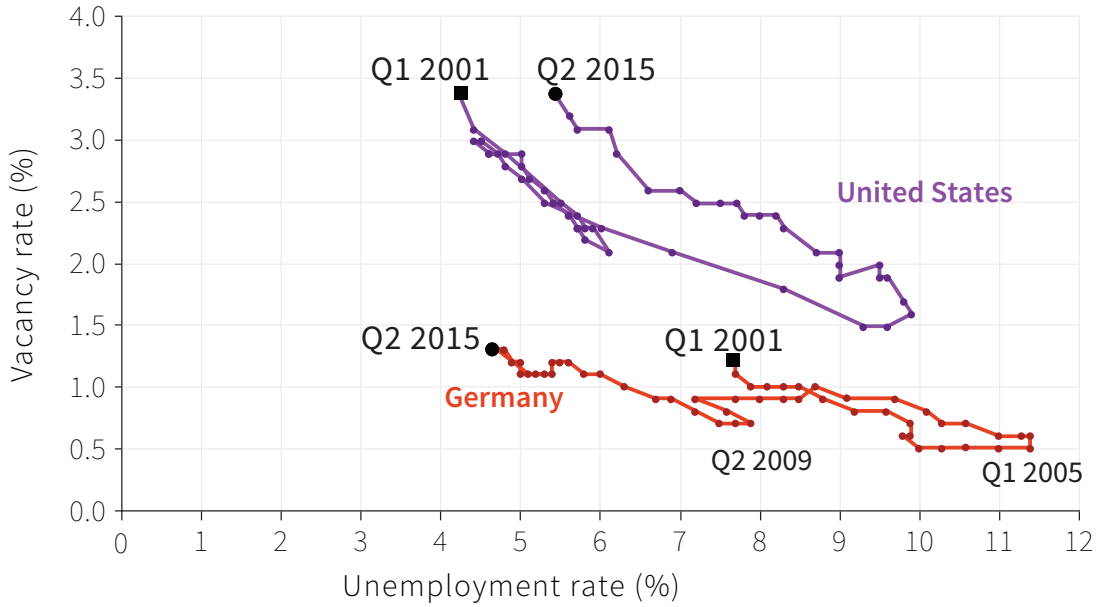


Figure 15.7 *Beveridge curves for the US and Germany (Q1 2001 – Q2 2015).*

Source: *OECD Employment Outlook and OECD Labour Force Statistics: OECD. 2015. 'OECD Statistics.'*

Why are there vacant jobs that are not filled, and unemployed people looking for a job at the same time? We can think of matching being tricky in many parts of life. For example, think of our love lives: how often are we looking for the perfect partner but are unable to find someone suitable?

What prevents newly unemployed people from being matched with newly posted jobs (we call this process *labour market matching*)?

- *A mismatch between the location and nature of the workers looking for jobs and the jobs looking for workers:* This is sometimes a matter of skills required by firms and the skills of jobseekers. For example, This research explains that one of the reasons for inefficiency in the US labour market in recent years has been that vacancies are concentrated in a few industries. The telephone engineer whose job was recently eliminated may not have the computer skills required to fill the vacancies in the company's billing department. Or the redundant workers and the vacancies may be located in different parts of the country. Travelling to another area to find a job would mean severing ties with neighbours, schools and family.

- *Either jobseekers or those seeking to hire may not have relevant information:* As we have seen in Unit 6, economic actors with different skills and needs— jobseekers and firms in this example—look for opportunities for mutual gains from trade. But the firm and the jobseeker may not know about each other (although there is evidence that technology is improving this matching process).

Matching should be easier when there is a large pool of the unemployed from which to select. Observing a combination of high unemployment and a large number of vacancies is an indicator of inefficiency in the matching process in the labour market.

Notice three things about the German and American Beveridge curves shown in Figure 15.7:

- *Both curves slope downward, as expected:* The US data oscillates between vacancy rates of about 3% with unemployment rates between 3% and 4% (at the top of the business cycle) to vacancy rates of a little over 2% and unemployment around 6% (at the trough of the cycle).
- *The position of each nation's Beveridge curves is different:* The German labour market appears to do a better job of matching workers seeking jobs to jobs seeking workers. To see this, notice that the vacancy rate in Germany for every year is lower than in the US for any year, although the two countries experienced a common range of unemployment rates. So, fewer job openings were wasted in Germany.
- *Both the curves shifted over the course of the decade:* The German curve, having established itself over the period Q1 2001 to Q1 2005, turned towards the origin and established a new Beveridge curve in the period Q2 2009 to Q1 2012. The latter Beveridge curve was closer to the origin, with a smaller sum of the vacancy rate and the unemployment rate than before.

How did this improvement in the German labour market occur? New policies called the *Hartz reforms* seemed to have worked. Enacted between 2003 and 2005, the Hartz reforms provided more adequate guidance to unemployed workers in finding work, and reduced the level of unemployment benefits sooner, so as to provide the unemployed with a stronger motive to search.

The US curve shifted too but, unlike Germany, conditions deteriorated. For the period Q1 2001 to Q2 2009 the US seems as if it is moving along a curve. After that the curve moves out from the origin and then seems to establish a new curve, above and to the right of the older one, suggesting the American labour market became less efficient in matching workers to jobs. Between 2001 and 2008 business cycle movements displaced workers in all industries all over the country in the usual way, so there wasn't much of a geographical and skills mismatch between workers looking for work and vacant jobs, so why did the Beveridge curve move?

- *Many redundancies in one industry:* The global financial crisis between 2008 and 2009, and the recession that followed, particularly affected the housing construction industry. There was a skill-based mismatch between the unemployed and vacancies.
- *The collapse of US house prices:* When house prices fell, many homeowners were trapped in a house that was worth less than they had paid for it. They could not sell their house and move to an area with more job vacancies, and this restricted their choice of jobs.

The result was that the economy moved into a situation where, for a given level of vacancies, there was a higher rate of unemployment.

DISCUSS 15.4: BEVERIDGE CURVES AND THE GERMAN LABOUR MARKET

Note: Although according to the Beveridge curves, the German labour market does a better job at matching workers with job openings, average unemployment in Germany over the period in Figure 15.7 was higher than in the US.

Refer back to section 13.10 to suggest hypotheses that could account for this. You may also find section 9.3 and section 12.3 useful for your answer too. What kind of data could be used to find support for your hypotheses?

15.4 INVESTMENT, FIRM ENTRY AND THE PROFIT CURVE IN THE LONG RUN

In Figure 15.1 we saw the remarkable divergence in unemployment rates across advanced economies that began in the 1970s. In the most recent period shown on the chart, European countries like Spain, Greece or France experienced very high unemployment rates, from around 10% in France to more than 20% in Spain, while in other countries, especially the east Asian (South Korea, Japan) and north European ones (Austria, Norway, Netherlands, Switzerland and Germany), unemployment was between 5% and 6%.

To explain the main shifts over time, and differences among countries, in the unemployment rate illustrated in Figure 15.1, we extend concepts from earlier units to model the long run. In the long-run model, things that may change slowly and

which are assumed to be constant in medium- or short-run models—such as the size of the capital stock, and the firms operating in the economy—can fully adjust to a change in economic conditions.

Determinants of economic performance in the long run

In the *long run*, the unemployment rate will depend on how well a country's policies and institutions address the two big incentive problems of a capitalist economy.

- *Work incentives:* Wage and salary workers must work hard and well, even though it is difficult to design and enforce contracts that accomplish this (as we saw in Unit 6).
- *Investment incentives:* The owners of firms must invest in job creation when they could invest abroad—or simply use their profits to buy consumption goods, and not invest at all. As we saw in Unit 13, firms considering investment decisions will take account not only of the rate of profit after taxes but also the risk of adverse changes such as hostile legislation or even confiscation of their property, which is referred to as *expropriation risk*. Just as workers cannot be forced to work hard, but have to be motivated to do so, firms cannot be forced to create new jobs or keep open existing ones.

Solving both problems simultaneously would mean a low level of unemployment at the same time as rapidly rising wages. But ways of addressing one of these problems may make it difficult to address the other. Policies that lead to very high wages may induce employees to work hard, for example, but leave owners of firms with little incentive to invest in creating new productive capacity.

In the next section we will see that countries differ in how successfully they address these two incentive problems simultaneously.

The *wage curve* that we have used in Units 6, 9, 13 and 14 shows that wages must be higher when unemployed workers expect to easily find a new job, or when they receive a generous unemployment benefit, both of which reduce the cost of job loss. This is why the wage curve rises when the employment level increases, and why an increase in the unemployment benefit will shift the curve upward, as this research demonstrates.

The necessary incentives for investment by owners of firms (Unit 14) are represented by the *profit curve* in the labour market model.

We will extend the labour market model to the long run by allowing firms to enter and exit and owners to expand the capital stock or allow it to shrink. To simplify, let's assume that firms are all of a given size, and that the capital stock grows or shrinks simply by the addition or subtraction of firms.

We define the *long-run equilibrium in the labour market* as a situation in which not only real wages and the employment level, but also the number of firms, is constant (remember *equilibrium* is always defined by what is unchanging, unless there is some force for change from things not considered in the model).

There are two conditions that determine how the number of firms may change:

- *Firm exit due to a low markup*: Owners may withdraw their funds or even close firms if the existing markup is too low, meaning that the expected rate of profit after taxes is not attractive relative to the alternative uses to which the owners could put their assets. These alternative uses could be investing in foreign subsidiaries, or outsourcing part of the production process, or buying government bonds, or distributing its profits as dividends to the owners. In this case the number of firms falls.
- *Firm entry due to a high markup:* If the markup is sufficiently high, the resulting high profit rate will attract new firms to enter the economy.

When is firm exit due to too low a markup likely to happen? This will occur when the economy is highly competitive as a result of a great number of competing firms, resulting in a high elasticity of demand for the firm's products and hence a small markup. When there are "too many" firms to sustain a high enough markup, then firms will leave, tending to raise the markup.

Similarly, when there are few firms in the economy, the degree of competition will be limited, the markup will be high, and the resulting profit rate will be sufficient to attract new firms to enter. As a result, the economy will become more competitive and the markup will fall.

This means that the markup has a tendency to self-correct: if it is too low then firms will exit and it will rise; if it is too high then firms will enter and it will decline.

Figure 15.8a illustrates this process by showing how the number of firms and the profit-maximising markup are related. For each number of firms, the downward-sloping line gives the markup that maximises the firm's profits. It slopes downward because:

- The more firms there are, the more competitive the economy is.
- This means a higher elasticity of demand facing the firms when they sell their products (less "steep" demand curves).
- The markup that maximises the firm's profits will fall, because, as we saw in Unit 7, the markup, μ, is *1/(elasticity of demand)*.

The other line in the figure is horizontal and shows the markup that is just sufficient to retain the existing number of firms, which we call μ^*. Consider what would happen if the number of firms, n, in the economy were 190. The economy is at B and the markup would exceed μ^* so new firms would enter, attracted by high profits. The opposite would occur if the number of firms were 250, with firms exiting due to low profits. This is why the number of firms will be stable at 210.

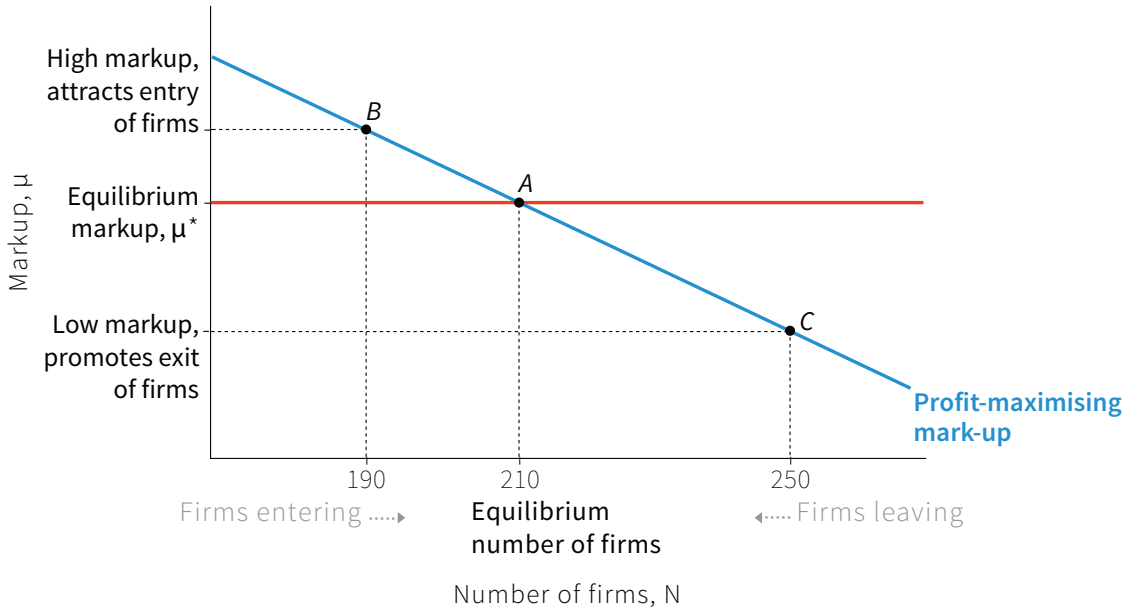


Figure 15.8a *Firm entry, exit and the equilibrium markup.*

Now, using Figure 15.8a think what would occur if as a result of a change of government, the risk of expropriation of private property by the government decreased. This could include changes in legislation that reduce the probability that the government will take over firms or implement unpredictable changes in taxation. This would have the following effects:

- *The equilibrium markup μ^* falls:* With the fall in risk, it takes less profits to attract new firms.
- *Firms enter as a result:* The economy grows.

This is shown in Figure 15.8b.

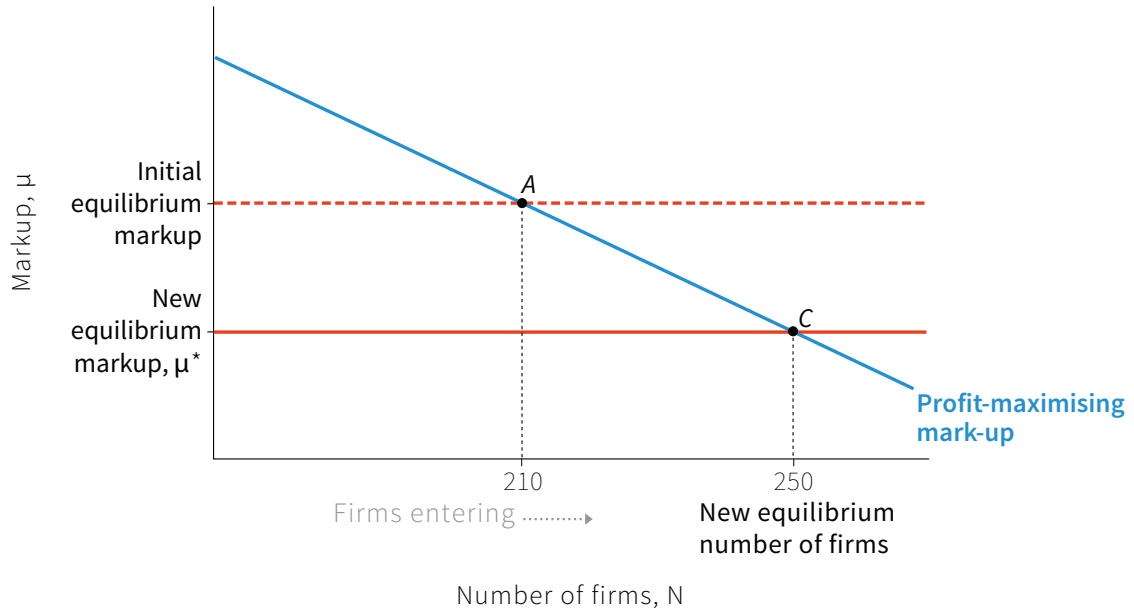


Figure 15.8b *An improvement in conditions for doing business: Firm entry, exit and the long-run markup.*

From the equilibrium markup to the profit curve in the long run

As before, once we know the markup μ^* and the productivity of labour λ, we know the real wage w that must result: it is the share of output per worker that is not claimed by the employer through the markup. The long-run profit curve is given by:

$$w = \lambda(1 - \mu^*)$$

This allows us to translate the equilibrium markup into the level of the wage, which fixes the height of the profit curve: see Figure 15.9. In the left-hand panel, the equation of the long-run profit curve is drawn with the equilibrium markup on the horizontal axis and the wage on the vertical axis: with a zero markup, the wage is equal to output per worker; and when the markup is equal to 1 (or equivalently 100%), the wage is equal to zero.

The right-hand panel of Figure 15.9 shows the long-run profit curve at different levels of the long-run equilibrium markup). We can summarise the factors that will shift the long run profit curve through their effects on either output per worker or the markup.

The long-run profit curve is higher:

- The higher the output per worker
- The lower the long-run markup at which entry and exit are zero

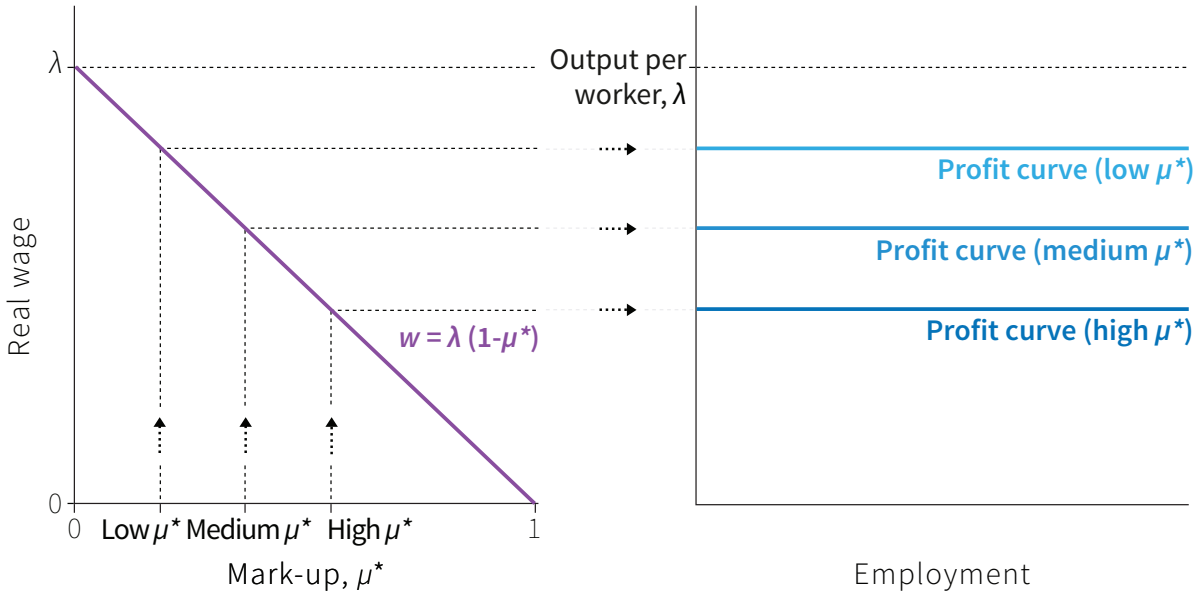


Figure 15.9 *Changes in the long run markup shift the profit curve.*

What lowers the markup at which entry and exit are zero?

- Higher competition
- Lower risk of expropriation of owners at home
- Higher quality environment for doing business: for example, human capital or infrastructure
- Lower expected long-run tax rate
- Lower opportunity cost of capital: for example, a lower interest rate on bonds
- Higher risk of expropriation of owners abroad
- Lower expected long-term cost of imported materials

LONG-RUN PROFIT CURVE

Once we know the equilibrium markup μ^* and the productivity of labour λ we know the real wage w is given by:

$$w = \lambda(1 - \mu^*)$$

- w is the share of output per worker that is not claimed by the employer through the markup.

DISCUSS 15.5: MEASURING THE CONDITIONS FOR INVESTMENT

Go to the World Bank's *Doing Business* database.

1. For 20 countries of your choice, collect (download) data on three characteristics from the *TOPICS* section that will affect the long-run markup. Justify your choices.

Now go to the *World DataBank* database.

2. Download GDP per capita data for the 20 countries of your choice. For each characteristic, create a scatterplot of the characteristic against GDP per capita and summarise the results. (Put the characteristic of the business environment on the horizontal axis and GDP per capita on the vertical axis.)
3. Provide reasons why a good business environment may raise GDP per capita.
4. Why might high GDP per capita improve the business environment?
5. From your answer, summarise potential challenges when interpreting the scatterplots.

15.5 NEW TECHNOLOGY, WAGES, AND UNEMPLOYMENT IN THE LONG RUN

We have seen that, contrary to the fears of the Luddites, the constant increase in the amount produced in an hour of work has not resulted in ever-increasing unemployment. Instead, technological progress—by roughly doubling the productivity of labour each generation in many countries—allowed a dramatic rise in the real wage consistent with firms making sufficient profits relative to the alternatives.

This upward shift in the profit curve is illustrated in Figure 15.10a, which shows the status quo ("Old technology") with the long-run equilibrium at *A*, and a technological advance that shifts the long-run equilibrium to *B*. At point *B*, the real wage is higher and so is the employment rate; unemployment is lower. The model shows that technological progress need not raise unemployment in the economy as a whole.

Before examining the experiences of unemployment in different countries, we need to understand:

- *What determines the rate of increase in the productivity of labour*? This accounts for the upward shift in the profit curve.
- *How does the economy shift from A to B*? Both are long-run equilibria in the labour market.
- *Why do we not see a constant decline in the unemployment rate*? If technological progress is constantly shifting the profit curve upward, surely the equilibrium shifts to higher and higher employment levels?

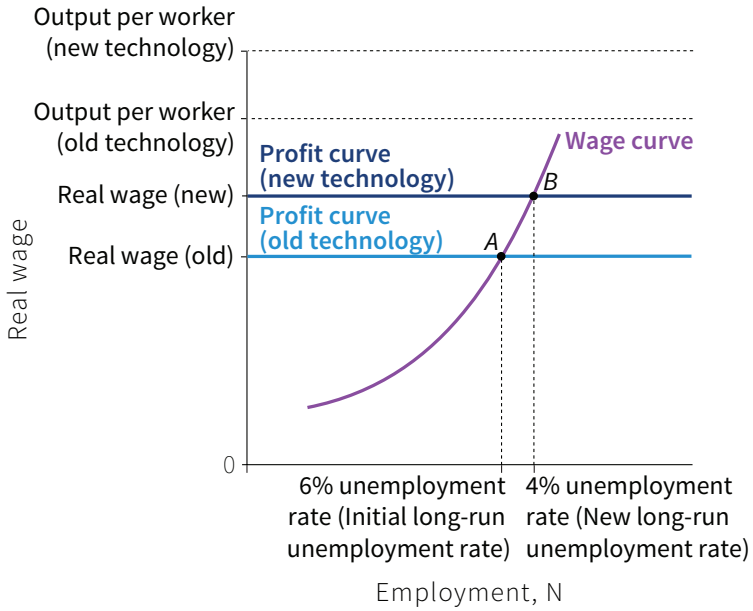


Figure 15.10a *The long-run unemployment rate and new technology.*

New knowledge and new technology: The diffusion gap

It often takes years if not decades before an improved technology is widely introduced in an economy. This *diffusion gap* causes differences between the productivity of labour in the most advanced firms compared to firms that lag technologically.

In the UK, one study found that the top firms are more than five times as productive as the bottom firms. Similar differences in productivity have been found in firms in India and China. In Indonesia's electronics industry—a part of the highly competitive global market—data from the late 1990s show that the firms at the 75th percentile were eight times as productive as those in the 25th percentile.

The low-productivity firms manage to stay in business because they pay lower wages to their employees, and in many cases earn a lower rate of profit on the owner's capital as well. Closing diffusion gaps can greatly increase the speed at which new knowledge and management practices are in widespread use.

This may occur when a union bargains for wages such that equivalent workers are paid the same throughout the economy. One consequence of this is that the least productive firms (which are also those paying low wages) will experience wage increases, making some of these firms unprofitable and putting them out of business. The union might also support government policies that complement its role in hastening the exit of unproductive firms, raising average productivity in the economy and shifting up the profit curve. In this case associations of workers can be part of the process of creative destruction rather than resisting it.

Associations of owners may also be part of the process of creative destruction by not seeking to prolong the life of unproductive firms, knowing that their demise is part of the process of making the pie larger.

But in many cases, employees and owners of the lagging firms do not act in this way. They gain protection through subsidies, tariff protection and bailouts that guarantee, at least for a time, the survival of the unproductive firm and its jobs.

The rate at which the economy's profit curve shifts upward depends on which of these attitudes towards the process of creative destruction is predominant. Economies differ greatly in this respect.

Adjustment to technological change

Economies differ too in how they make the journey from a *status quo* equilibrium like *A* to a new equilibrium such as *B* in Figure 15.10b.

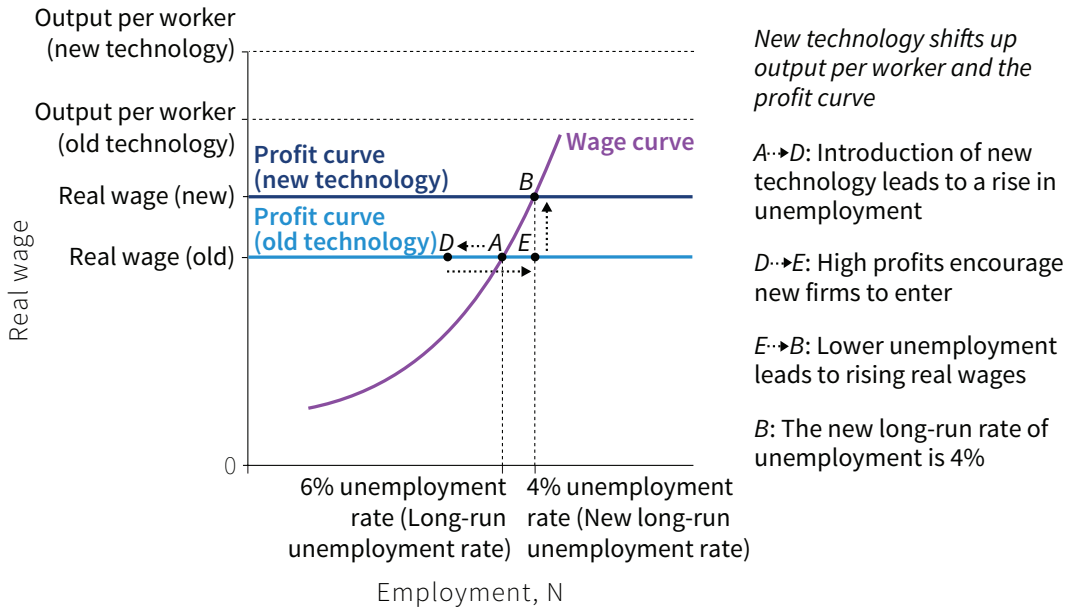


Figure 15.10b *The long-run unemployment rate and new technology.*

Recall that the profit curve in the long-run model is the level of the real wage such that firms will neither enter nor leave the economy. So the move from point A (at 6% unemployment) to point B (at 4% unemployment) occurred because firms entered the economy, a process that takes some time. What happened along the way? Here is a possible itinerary:

- *The implementation of the new technology displaces a number of workers from their jobs:* Looking at the whole economy, the increase in output per worker meant that to produce the same output, fewer workers were required. The immediate consequences would be firms closing and employment being cut back at

surviving firms, even as employment expanded in the firms introducing the new technology. This process might lead to the journey starting out with a move from *A* to *D* in Figure 15.10b, with the same wage and fewer jobs.

- *But economic profits are high at D*: The wage is well below the new profit curve. New firms will be attracted to the economy and investment will rise.
- *Unemployment eventually falls*: The economy moves from *D* to *E* in Figure 15.10b.
- *Firms have to set higher wages to secure adequate worker effort*: Unemployment is lower, so wages go up. Adjustment stops when the economy is at point *B* with higher real wages and lower long-run unemployment.

This looks like a win-win outcome. But it can take a long time, and along the way there will be losers.

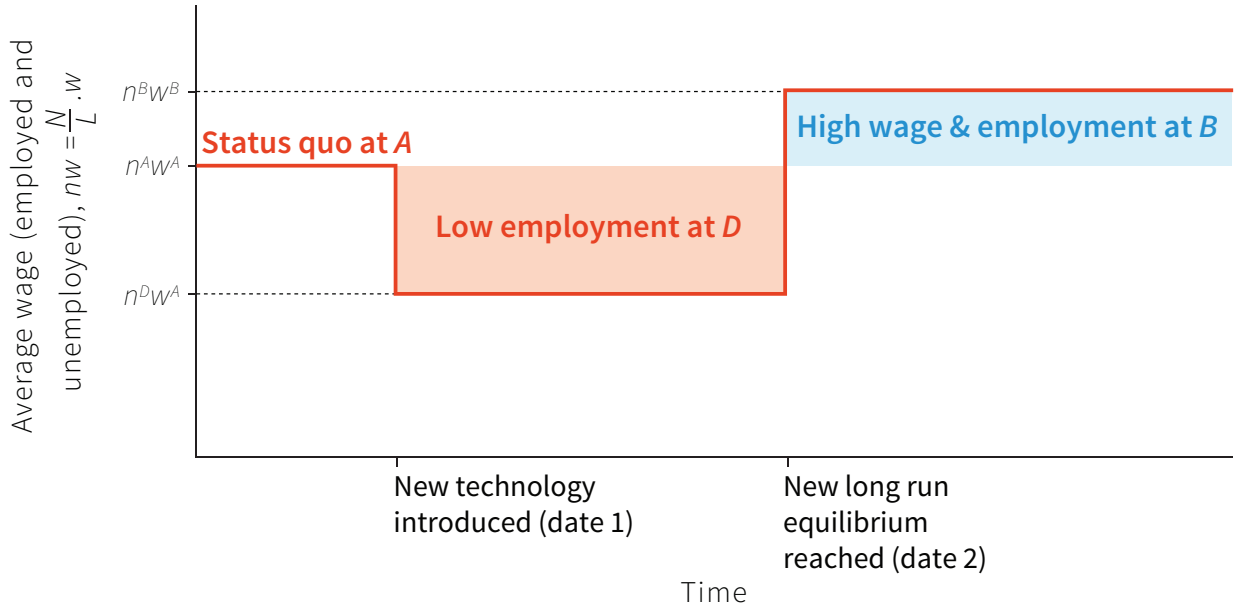


Figure 15.11 *Adjustment to technological change.*

Figure 15.11 considers the time path of adjustment for a simplified version of the journey from *A* to *B* just described. The vertical axis measures the average wage of a person working or seeking work, that is the fraction employed multiplied by the wage they receive (those not employed are assumed to receive no wage). If *N* is employment, *L* the labour force, and w the real wage, this is written as:

$$nw = \frac{N}{L} \times w$$

When new technology is introduced the fraction employed falls from n^A to n^D, the wage remaining unchanged. This situation persists until the economy moves (we assume, unrealistically, in one jump) to a new equilibrium with a higher wage and employment level.

Was this a win-win journey? Only if you look at the starting and end points or have a sufficiently long time horizon. The detour through greater unemployment costs workers an amount indicated by the pink shaded area, associated with low employment at *D*, which we call the employment *adjustment gap*. This will be offset by gains indicated in the open ended blue rectangle, where the wage and employment are higher at *B*.

The time between the introduction of new technology and the new long-run equilibrium would be measured in years or even decades, not weeks or months. Younger workers might have more to gain from the eventual higher wages and employment; older workers might never see the high wages in the *B* rectangle.

We have already seen that in Britain adjustment to the technological progress in the 18th and 19th centuries, called the Industrial Revolution, was not rapid. There was a prolonged delay before real wages began a sustained rise around 1830.

Just as was the case with the diffusion gap, public policies, trade union and employer association practices can alter the size of the employment adjustment gap.

Government policy can help in the reallocation of workers to new firms and sectors by providing job-matching and retraining services, and by providing generous but time-limited unemployment benefits. This helps workers released from failing firms to move quickly to better ones.

It also depends on institutions that could ease or hamper the creation of jobs in new sectors. If the wage is below the profit curve, profits are sufficient to create new investment and form new firms. This is part of the process of adjusting to creative destruction. Some countries have well-designed product market regulation and competition policy to make it easier to start a new business. In others incumbent businesses have succeeded in making it difficult for new firms to enter, slowing or preventing the economy moving to point *B*.

Fair-share bargaining and the upward shift of the wage curve

We saw in the data that persistent technological progress does not imply that the unemployment rate will fall over time (as at point B in Figures 15.11 and 15.10b).

An increase in labour productivity shifts up the profit curve. Without a wage increase, this would increase the profit rate and the share of the total output claimed by employers rather than workers. Firms would enter and unemployment would be lower. But the process of technological change may indirectly *bring about an upward shift in the wage curve* through some combination of the following mechanisms:

- *Unions bargain for restoring their previous share:* They argue that it is fair to claim their share of the economy's output.
- *Elected members of government may adopt more generous unemployment insurance:* As the economy adjusts to the new technology, they wish to assist those out of work.

- *Employees expect a greater disutility of effort on the job as a consequence of the new technology:* They resent their employers if wages do not rise. Employers must pay more to induce them to work.
- *Technological improvements in the countryside and urban migration associated with the implementation of new technology in manufacturing may raise rural incomes:* This would occur in developing countries with large rural sectors. Like a rise in unemployment benefits, this can be seen as an improvement in the *reservation option* for urban workers, and make the loss of an urban manufacturing job seem less of a deterrent. Urban employers must pay more as an inducement to work.

These consequences of an upward shift in the profit curve cause an upward shift in the wage curve, shown in Figure 15.12.

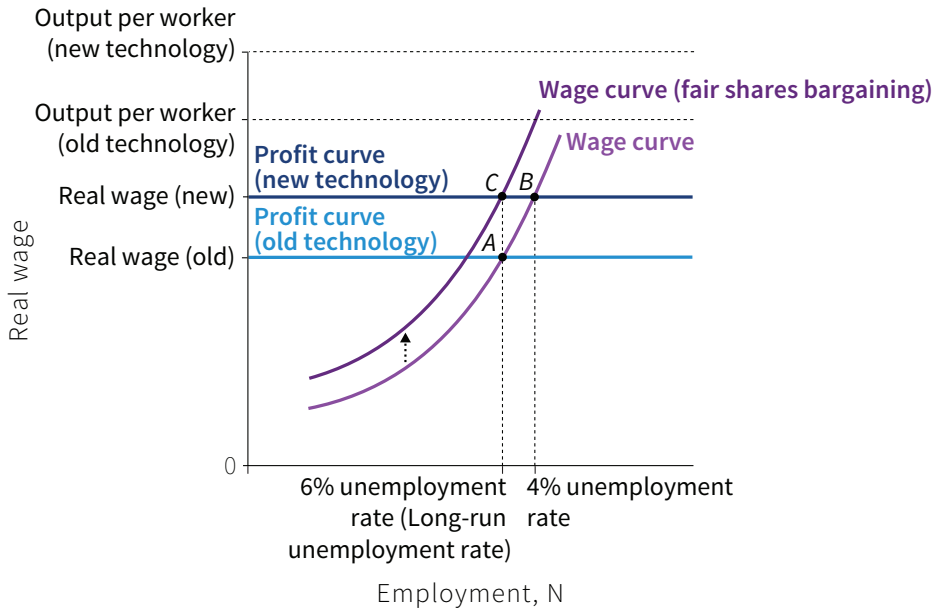


Figure 15.12 *The long-run unemployment rate and new technology: Fair-shares bargaining.*

Lessons from creative destruction and consumption smoothing

By this time, you may have noticed two recurring themes in this course:

- *Creative destruction:* Improvements in living standards often occur by a process of technological progress in which jobs, skills, entire sectors and communities become obsolete and are abandoned. We study this process in Units 1, 2, 15 and 20.
- *Consumption smoothing:* Households faced with shocks to their income seek to even out the ups and downs of their standard of living by smoothing their consumption through borrowing, unemployment insurance, mutual assistance among family and friends and other forms of co-insurance. We study this process in Units 11 to 13.

The two are related: people suffering from destruction will suffer less if they can smooth consumption. Economies differ greatly in the extent to which their policies, culture and institutions allow consumption smoothing. In those that do this well, resistance to the creative-destructive forces of technological progress is likely to be low. In those that do not, owners and employees alike will try to find ways to resist (or halt) the process of creative destruction, preferring to defend their firm's assets and existing jobs.

The attitude to the process of job destruction and creation is an example. In countries with adequate consumption-smoothing opportunities, trade unions tend not to insist on a worker's right to keep a particular job. Instead they demand adequate new job opportunities, and support in searching and training for new work.

In other countries unions and government policy seek to protect the *status quo* matching of workers to jobs, for example by making it more difficult to terminate a labour contract, even when the worker has performed inadequately. This *employment protection legislation* may be harmful to labour market performance by enlarging the diffusion and adjustment gaps, and retarding the rate of technical progress—and at the same time pushing the wage curve up.

These differing responses to the opportunities and challenges presented by creative destruction will help us understand why some economies performed better than others in recent history.

15.6: INSTITUTIONS AND POLICIES: WHY DO SOME COUNTRIES DO BETTER THAN OTHERS?

What do we mean by good performance and a good outcome? The answer matters because citizens who vote for parties with alternative economic programs, and policymakers who attempt to improve those programs, will need some concept of what is desirable—either for the individual, the policymaker or the nation.

As we saw in Unit 3, people value their free time as well as their access to goods. We should include their reward per hour of work in our evaluation of outcomes. In any given year, a good performance is one in which unemployment is low and real wages per hour are high. Putting this into a dynamic setting, and evaluating an economy over many years, we judge performance as good if a country combines rapid growth of real wages per worker hour with low unemployment.

There are of course other dimensions of performance of the economy in the long run that most people care about. We may care whether the distribution of economic rewards is fair, whether the economy's relationship with the natural environment

is sustainable, or about the extent to which households are subjected to economic insecurity through business cycle fluctuations. But here we focus solely on the growth in real wages per hour and the unemployment rate.

We use the labour market model and the Beveridge curve to see that achieving good performance requires an economy to have two capacities:

- *To raise the profit curve and restrain the upward shift of the wage curve:* So that both hourly wage growth and the long-run employment rate are high
- *To adjust rapidly and fully:* So that the economy can take advantage of opportunities from technological change.

Technological change means jobs disappearing in firms in which new technology substitutes for workers. Jobs also disappear as new firms enter and those unable to adapt to the new conditions shut down. The Beveridge curve highlights the importance of matching workers and vacancies in the labour market. In Figure 15.10b, we saw that the impact of new technology is initially to displace workers: the Beveridge curve summarises the economy's ability to rapidly redeploy displaced workers, shortening the period the economy spends around Figure 15.10b's point *D*.

Figure 15.13 shows long-run performance (over a 40-year period) for a group of advanced economies. It uses the criteria of real wage growth and unemployment rates. We study a long period because we do not want our evaluation of long-term performance to be affected by the particular phase of the business cycle in which a country finds itself (it will look much better at the peak than at the trough). We use wages in manufacturing because they are measured in ways that are more accurately comparable across nations—although this is not ideal, because a shrinking share of employment is in manufacturing and the share varies across countries.

Good performance places a country in the top-left corner of Figure 15.13 with high wage growth and low unemployment; bad performance places a country in the bottom-right corner. Since we value both high wage growth and low unemployment, we may be prepared to tolerate low wage growth if it is associated with a lower level of unemployment. This means that we can represent a citizen's indifference curve as a ray from the origin. Steeper rays are better, and a country's performance is measured by the steepness of a ray from the origin to that country's observation. If you look at Figure 15.13, and take Belgium (BEL) as an example: a Belgian citizen would prefer to be on a steeper ray, like that of Germany (GER), with lower unemployment and higher wage growth.

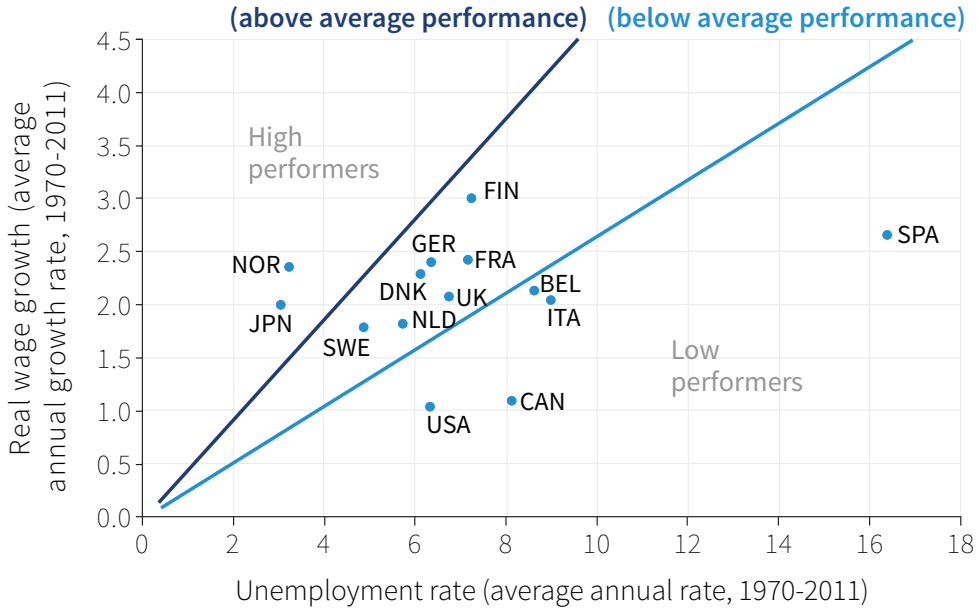


Figure 15.13 *Long-run unemployment and real wage growth across the OECD.*

Source: *OECD. 2015. 'OECD Statistics.'; BLS data for Spanish real wages available only from 1979. Spanish real wage growth for 1970-1979 has therefore been estimated using Tables 15.25 and 16.5 from Barciela López, Carlos, Albert Carreras, and Xavier Tafunell. 2005. Estadísticas Históricas de España: Siglos XIX-XX. Bilbao: Fundación BBVA.*

DISCUSS 15.6: YOU ARE THE POLICYMAKER

1. If as a citizen or policymaker, you cared only about wage growth, what would your indifference curves look like?
2. According to Figure 15.13, which countries would be the best performers and which would be the worst?
3. If you cared only about the unemployment rate, what would your indifference curves look like? Which countries would be the stars and which would be the laggards in this case?
4. Draw an indifference curve based on your own personal preferences and justify your choice.
5. Suppose you were an employee in the manufacturing sector in one of the countries in Figure 15.13. Knowing their histories of wage growth and unemployment, which country would you choose to live in and why?
6. Rank your top and bottom three choices of country. Explain the reasons for your ranking.

(For your ranking, assume that current wages are the latest figures given in this report. For simplicity, assume that for this question the unemployment insurance is 50% of the wage and the number of weeks without work will be numerically equal to the unemployment rate. For example, in the US: the unemployment rate is 6%, so you would expect to be unemployed six weeks in a year.)

The two rays in Figure 15.13 divide the countries into three groups. The best performers over the 40-year period from 1970 to 2011 are Norway and Japan. The worst are Belgium, Italy, US, Canada and Spain. The poor performance of the US is in part due to the fact that it started with higher wages in 1970, because it was the world's technology leader during this period, as we saw in Figure 15.4. This meant that other nations could learn from it, rapidly raising their productivity. Similar arguments apply to Canada. For this reason we do not take these two countries as representative of the poor performers, although real wages have grown much more slowly than productivity in the US, so most US citizens did not benefit very much from economic growth in this period. Notice that some of the best performers (on steeper rays from the origin) like Norway, Finland, Sweden and Germany have powerful unions, and the Nordic countries (including Denmark) have some of the most generous unemployment benefits in the world.

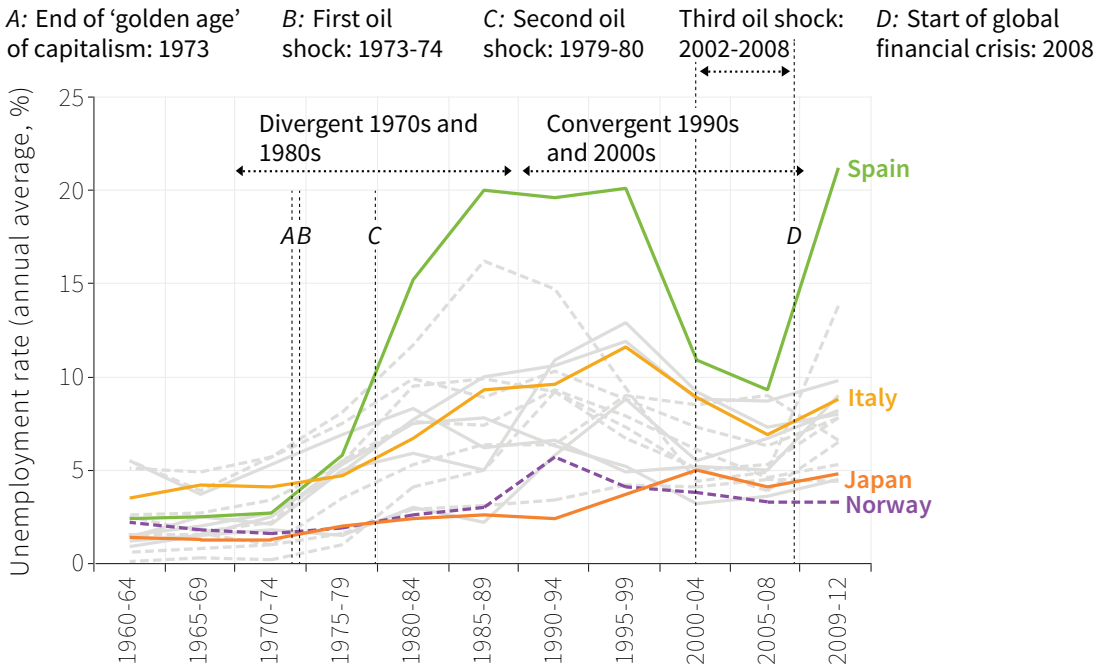


Figure 15.14 *Unemployment rate for stars and laggards (1960-2012).*

Source: *Data from 1960-2004: Howell, David R, Dean Baker, Andrew Glyn, and John Schmitt. 2007. 'Are Protective Labor Market Institutions at the Root of Unemployment? A Critical Review of the Evidence.' Capitalism and Society 2 (1); Data from 2005 to 2012: OECD harmonized unemployment rates, OECD. 2015. 'OECD Statistics.'*

Figure 15.14 reproduces the same unemployment data from Figure 15.1, but with some of the stars and the laggards highlighted.

We shall see that the model in this unit provides a useful framework for understanding the stars and laggards of the labour market. We will now show how to use the model (by shifting the profit and wage curves) to explain the way institutions and policies affect real wage growth and unemployment in the long run.

15.7 TECHNOLOGICAL CHANGE, LABOUR MARKETS, AND TRADE UNIONS

Policies and institutions make a difference. The models shed light on the experience of some of the best and worst performers. We take three countries: Norway and Japan as good performers, and Spain as a poor performer.

In Norway and Spain, but not in Japan, unions are important. In Norway, more than half of all wage and salary workers are trade union members and union wage deals affect most workers in the economy. In Spain, although union wage deals are important for the entire economy, less than one-fifth of Spanish workers are in unions.

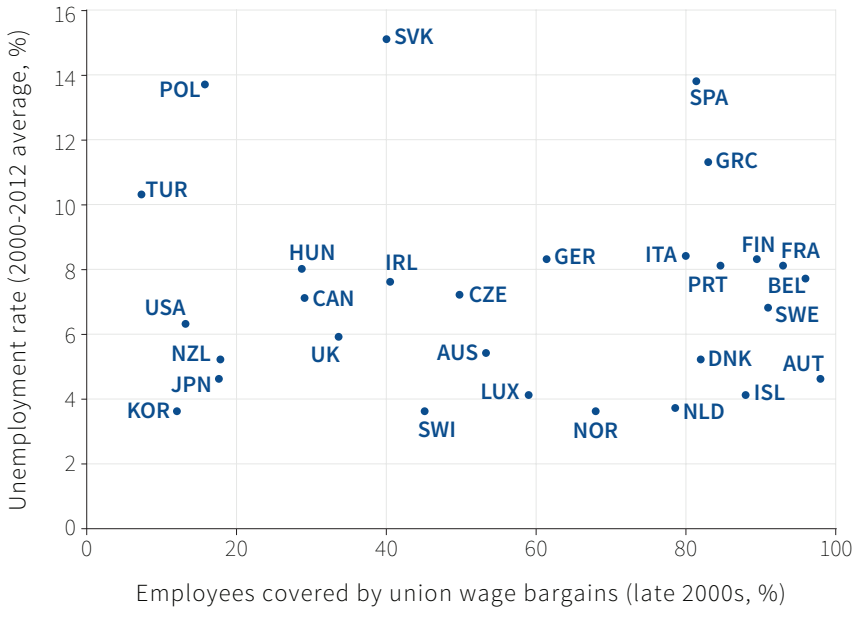


Figure 15.15 *Union wage bargaining coverage and unemployment across the OECD (2000-2012).*

Source: *OECD. 2015. 'OECD Statistics.'. Labour force statistics. Visser, Jelle. 2013. 'ICTWSS: Database on Institutional Characteristics of Trade Unions, Wage Setting, State Intervention and Social Pacts in 51 Countries between 1960 and 2014.' Amsterdam Institute for Advanced Labour Studies (AIAS).*

Figure 15.15 provides information on the importance of union wage deals and unemployment. On the horizontal axis we plot the percentage of employees whose wages are determined by union wage deals. As you can see, in some European countries, union wage deals cover almost all employees. And in the set of countries with coverage of more than 80%, unemployment rates range from less than 4% (Netherlands) to almost 14% (Spain). Figure 15.15 suggests that there is no tendency for unemployment to be higher in countries in which unions are more influential in wage-setting. Low unemployment is found in countries extending across the whole range of union strength: compare South Korea and the Netherlands; or Japan and Austria; or the US and Sweden.

Just as the employer does not offer the lowest wage possible, most unions do not seek the highest wage they could win in bargaining. Employers offer wages above the minimum because they cannot control how hard the worker works. Unions do not bargain for the maximum wage possible (the wage that would leave none of the pie for the owners) because unions cannot control the firm's decisions about hiring, firing, and investment, and higher wages may reduce employment by reducing the firm's profits.

A union organised across many firms and sectors will not exploit all the bargaining power it possesses. It knows that large wage gains will lead to:

- *In the medium run:* Restrictive aggregate demand policies, as the government and central bank seek to keep inflation close to target (as we saw in Unit 14).
- *In the long run:* The exit of firms and a smaller stock of capital goods, which will slow the rate of productivity growth.

Unions that act this way are called *inclusive unions*. Non-inclusive unions may bargain for high wages in their own corner of the economy without regard for the effects on other firms, and workers, both employed and unemployed. Employers' associations that take account of the interests of all businesses, including those that might enter an industry and compete with its incumbent firms, are called *inclusive business or employers' associations*.

The Nordic case: Inclusive unions and employers' associations

This inclusive behaviour is exactly what the trade unions and employers' associations of Norway (as well as in the other Nordic countries) did over this period: their centralised wage bargaining insisted on a common wage for a given kind of labour, depriving low-productivity firms of access to inexpensive labour and driving many of them out of business. As workers were quickly redeployed to employment in more productive firms, the main impact was to raise average labour productivity, pushing up the profit curve and allowing higher wages.

Inclusive trade unions also support generous income floors and high quality publicly provided health, occupational retraining and educational services: all of which reduce the risk to which most individuals are exposed. This has the effect of making the creative destruction of technological change less destructive of people's personal lives, and allows them to be generally more open to change and to risk taking, both attributes essential for a technologically dynamic society. These so-called "active labour market policies" improve the matching process between workers looking for work and job vacancies looking for workers. A result is that workers whose jobs are eliminated (for example by the failure of low-productivity firms under the pressure of centrally bargained uniform wages) more quickly find an alternative job. A result is a Beveridge curve close to the origin, superior to both the German and US Beveridge curves (shown in Figure 15.7). It is far inside that of Spain, as we see in Figure 15.16.

Figure 15.16 *Beveridge curves for Spain and Norway (Q1 2001 – Q4 2013).*

Source: *OECD Employment Outlook: OECD. 2015. 'OECD Statistics.'*

An inclusive union knows that the economy has to respect the two major incentive problems of a capitalist economy: providing incentives for workers to work and for employers to invest. In some cases—for example in Sweden with its highly centralised trade union federation—trade union leaders knew and persuaded their members that, in the long run, pushing down the wage curve will increase employment; and it will not, in the long run, reduce wages.

As a result the inclusive unions of the Nordic countries—Norway, Sweden, Finland and Denmark—set their wage demands in accordance with the productivity of labour. When it rose they demanded a fair share. They had bargaining power from low unemployment, high membership and their ability to implement wage agreements across the economy, but they did not use this power to push the wage

curve up unless this was warranted by productivity growth. These unions also supported quality of legislation and policies that make working less onerous, shifting downward the wage curve, and further expanding long-run employment.

The Japanese case: Inclusive employers' associations

In contrast to the Nordic countries, Japanese unions are weak, but workers are well organised in the large companies. Employers' associations are strong and work to coordinate wage setting among the large firms. These associations therefore operate in a similar way to the unions in Norway: the impact of wage decisions on the economy as a whole is taken into account when wages are set. Specifically, the corporations deliberately do not compete in hiring workers, so as to avoid raising wages.

The Spanish case: Non-inclusive unions

Unions protect jobs in Spain, supported by government policy. Wage-setters in Spain are strong enough to wield power, but are not inclusive. A combination of non-inclusive unions and supporting government legislation to protect jobs may help to account for the poor performance of the Spanish labour market.

Based on the model, we would predict high unemployment in Spain, and low unemployment in Norway and Japan. And that is what we see in the data.

Unemployment benefits and unemployment

The employment-enhancing effects of inclusive trade union and government co-insurance policies may help to explain an apparent anomaly: countries with generous unemployment benefits do not have higher rates of unemployment (see Figure 15.17).

This is anomalous, because in our model an increase in the unemployment benefit would, *ceteris paribus*, reduce the workers' cost of job loss, and shift the wage curve up.

The contrast between unemployment rates and benefits in Norway and Italy illustrates the point. An unemployed person gets a benefit of almost 50% of previous gross earnings in Norway, and unemployment is low; by contrast, benefits in Italy offer a 10% gross replacement rate, and unemployment is much higher than in Norway. The implication is that countries that are able to implement generous but well-designed unemployment insurance schemes, coordinated with job placement services and other active labour market policies, can achieve low rates of unemployment. You can read more about the role of institutions in European unemployment in these papers.

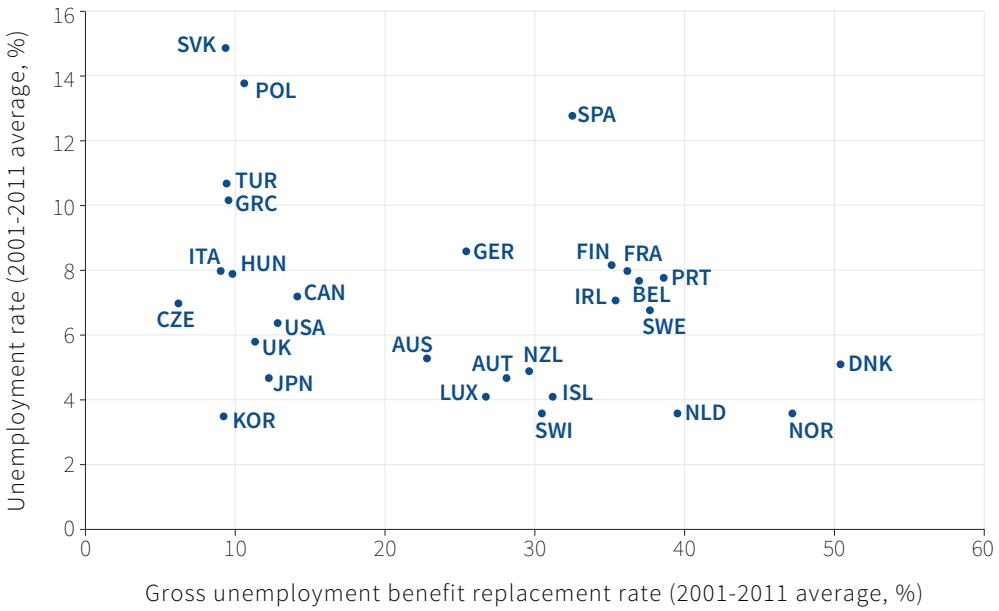


Figure 15.17 *Unemployment benefit generosity and unemployment across the OECD (2001-2011).*

Source: *OECD. 2015. 'OECD Statistics.'*

DISCUSS 15.7: LABOUR FORCE STATISTICS AND TAX-BENEFIT MODELS

Some people have argued that high unemployment in some European countries relative to the US during the 1990s and 2000s was due to the existence of rigid labour-market institutions (defined as including powerful unions, generous unemployment benefits and strong employment protection legislation).

1. Using Figure 15.1, check if the unemployment rate has always been higher in most European countries compared to the US.
2. From what you have learned from this section, and by looking at Figures 15.1, 15.15 and 15.17, discuss the claim that high unemployment in Europe was due to the existence of rigid labour market institutions

15.8 CHANGES IN INSTITUTIONS AND TECHNOLOGY

We have seen that differences in institutions make a big difference for employment and wage growth, and that citizens of Spain might wish to have institutions like those of Japan or a Nordic country. But changing institutions is difficult because it inevitably involves creating winners and losers. Yet some economies have succeeded in creating institutions that work relatively well, at least by the standards we have set. Germany and Finland are examples. As we will see, the same is true of the UK and the Netherlands.

Countries that changed their policies changed their fortunes. Both the UK and the Netherlands suffered sharply increased unemployment rates in the 1970s and early 1980s, due to the first and second oil shocks—which shifted the profit curve down—and the increased bargaining power of labour, which caused an upward shift in the wage curve. This is illustrated in Figure 15.18. But a change in policy eventually turned the bad news around. In the UK, the unemployment rate fell from 11.6% in 1985 to 5.1% in 2002; in the Netherlands it fell from 9.2% to 2.8% over the same period.

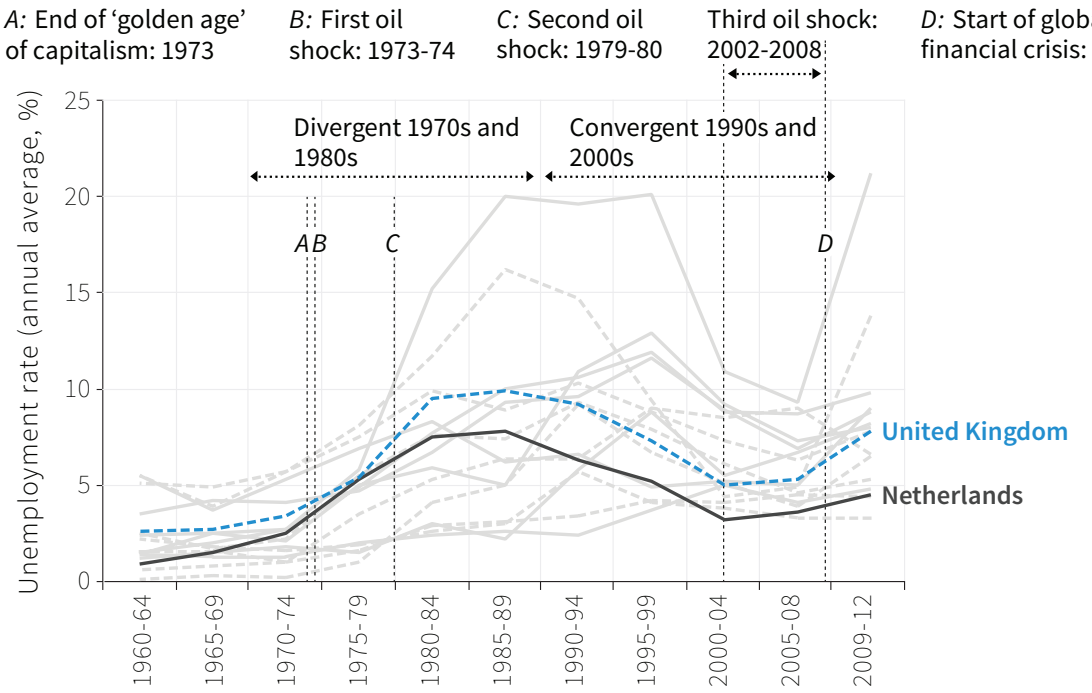


Figure 15.18 *Different ways of pushing down the wage curve: The Netherlands and the UK.*

Source: *Howell, David R, Dean Baker, Andrew Glyn, and John Schmitt. 2007. 'Are Protective Labor Market Institutions at the Root of Unemployment? A Critical Review of the Evidence.' Capitalism and Society 2 (1); Data from 2005 to 2012: OECD harmonised unemployment rates: OECD. 2015. 'OECD Statistics.'*

Both countries turned around their economies and shifted the wage curves down, but they used different institutions and policies:

- *In the Dutch case:* Institutions became more inclusive, moving in a Nordic direction by common agreement.
- *In the British case:* Policy reduced the power of the non-inclusive unions and increased competition in labour markets.

In the Netherlands, a key component was an agreement in 1982 between employers and unions called the *Wassenaar Accord*. Unions offered wage restraint (a downward shift in the wage curve) and in exchange, the employers agreed to a reduction in working hours. The union agreed that the reduction in working hours would not increase labour costs (and hence would not shift the profit curve down). In the Dutch case, unions and employers' associations were capable of coordinating wage setting to achieve a better macroeconomic outcome. They were sufficiently powerful that they could ensure their members stuck to the agreement. The unions were exercising bargaining restraint in the interests of improved performance of the labour market, and hence in the economy as a whole.

In the UK, the wage curve also shifted down but, in this case, it was because of a fall in union power brought about by government policies designed to weaken the ability of the non-inclusive unions to organise strike action by changing industrial relations legislation.

The success of the US: The profit curve shifts up

The US was the only country that improved its performance on both the real wage growth and unemployment measures between the first and the second half of the 1970-2011 period, as illustrated in Figure 15.19.

When we evaluated labour market performance over the 40-year period from 1970 in Figure 15.13, the US looked a poor performer. The ray from the origin to the US was fairly flat. A partial explanation is that other countries were catching up to the US, and as a result, had stronger wage growth.

In the two decades from 1970-1990, real hourly wages in manufacturing grew 0.9% per annum, increasing to 1.2% per annum over the following 20 years. Unemployment fell from 6.8% in the first period to 5.7% in the second period. Although these are small improvements, the performance of most other countries deteriorated. The US joined the group of mid-performing economies in the 1991-2011 period, which included Japan, the Netherlands, Denmark, Finland, Germany, the UK and Sweden as shown in Figure 15.19. Note the large deterioration in performance of Germany, Italy and Spain (and other economies) between the first and second periods.

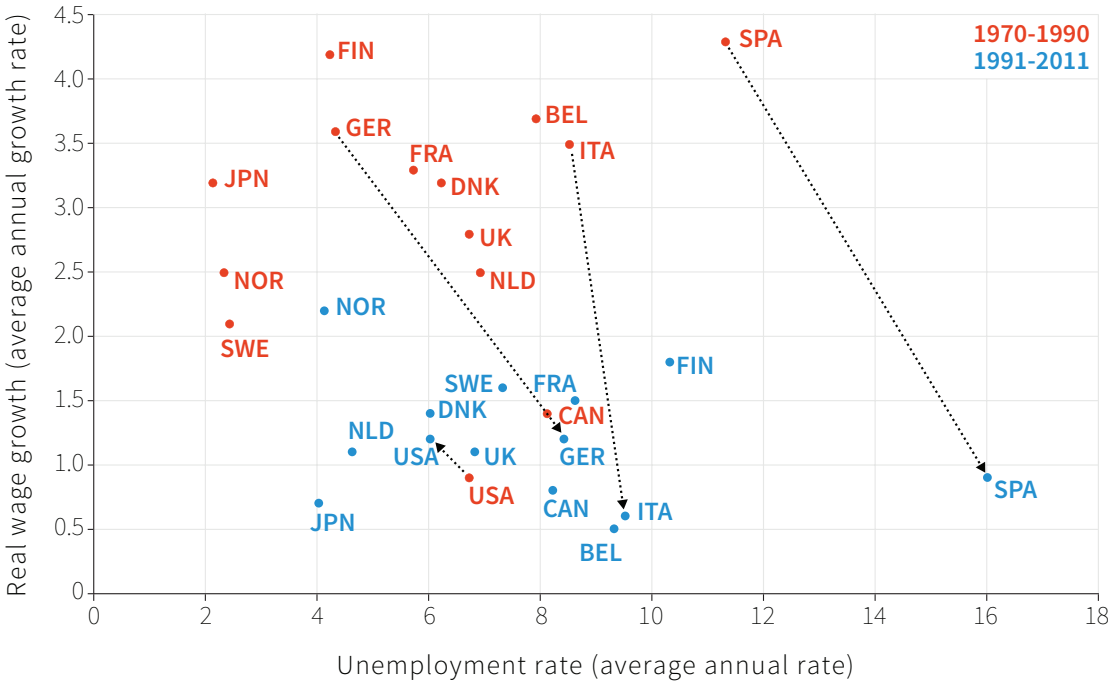


Figure 15.19 *Long-run unemployment and real wage growth in manufacturing across the OECD: Comparing 1970-1990 with 1991-2011.*

Source: *OECD. 2015. 'OECD Statistics.' OECD harmonised unemployment rates.*

An upward shift in the profit curve is necessary for a rise in real wages. This will also be associated with a fall in unemployment, *ceteris paribus*. Why did this upward shift occur in the US in the 1990s?

In Units 12 and 14, we looked at the hi-tech boom and bust in the US. The hi-tech boom dates from the middle of the 1990s and is reflected in productivity growth in the economy, which increased from a rate of 1.6% per year from 1980-1995 to 2.7% per year from 1995 to 2004. In contrast, in the European Union productivity growth fell from 2.8% to 1.6% over this period. More detailed data show that the contribution of the knowledge economy, including the use of ICT (information and communications technology), accounted for the gap in dynamism between the US and Europe in the post-1995 period.

A second factor that pushed the profit curve up in the US in the second period is that, like other countries, the government reduced the corporate tax rate. With lower taxes to pay, the real wage consistent with firms making sufficient profits to sustain employment rose.

A third factor behind improved US performance in the second period is that unions—very much of the non-inclusive type—became weaker. From the model, we know that this will have had the effect of keeping the wage curve down and supporting a lower unemployment rate. It is also likely that weaker unions did not strongly resist the implementation of new technology, causing faster diffusion through the economy. The US was the pioneer in the use of monitoring equipment for raising productivity

in logistics in retail, wholesale and transportation industries, for example, and was able to raise productivity in retail, wholesale and finance using this new technology and the related organisational changes.

The study of improved US labour market performance highlights mechanisms important in some emerging economies. South Korea is a good example, where rapid productivity growth, partly fuelled by technologies introduced from more advanced economies, shifted the profit curve up during the period sustaining high wage increases and low unemployment (South Korea would rank among the star performers in Figure 15.13 but the lack of comparable data means it is not included in the figure).

DISCUSS 15.8: THE LABOUR MARKET MODEL

1. Explain how to use the labour market model (wage curve and profit curve) to show the changes in labour market performance of the UK, the Netherlands and the US over the last 40 years as discussed in this section. (Draw diagrams for each country separately.)
2. What other factors may have contributed to the performance of these countries?

15.9 CONCLUSION

In 1914 Henry Ford, the founder of the Ford Motor Company, surprised his employees and competitors: all workers, he announced, would be paid a minimum of $5 per day, compared to $2.34 previously. This increase would cost half the company's profit at the time, and many economists thought that Ford would soon go bankrupt.

Why did he do this? The previous year, on average 13,623 people worked for him at any time but, during that year, more than 50,000 had left: 8,490 had been fired, the rest had quit. The fact that so many had quit, and that he had had to fire so many others, meant that he was paying barely more than their reservation wage. At $2.34 per hour the cost of job loss was not sufficient to deter turnover or to inspire hard work.

It worked. Quits fell to less than one-tenth of their previous level and Ford fired exactly 27 workers in the year following the pay rise. Despite doubling workers' pay, Ford's profits rose, because Ford produced more output per worker per hour.

We have learned that national economies differ not only in how rapidly they adjust to the opportunities offered by technological change and other changes of circumstance; they also differ in the wages and employment that they can sustain in the long run.

Figure 15.20 summarises the determinants of the unemployment rate and the growth rate of real wages, and notes the units where these concepts are discussed.

Figure 15.21 builds on Figure 15.20 by showing the institutions and policies that can affect the growth of real wages and the unemployment rate.

CONCEPTS INTRODUCED IN UNIT 15

Before you move on, review these definitions:

- *Creative destruction*
- *Marginal product of capital*
- *Job creation, job destruction*
- *Diminishing marginal product of capital*
- *Beveridge curve*
- *Labour market matching*
- *Long-run profit curve*
- *Equilibrium markup*
- *Diffusion gap, Adjustment gap*

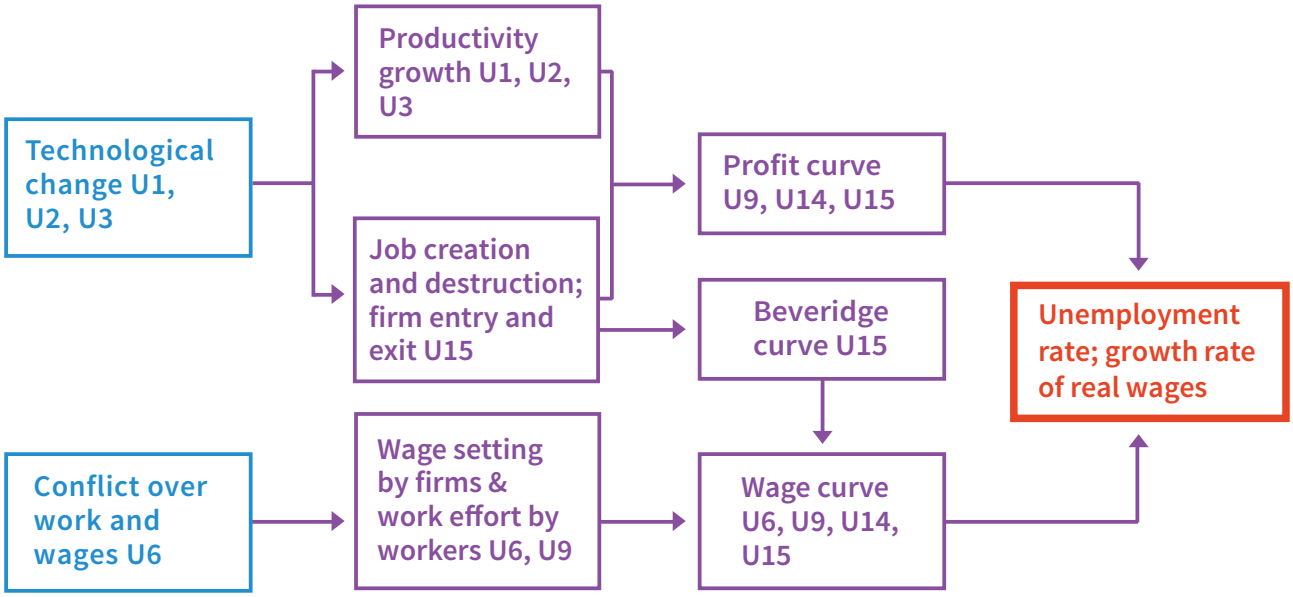


Figure 15.20 *The determinants of the unemployment rate and the growth rate of real wages.*

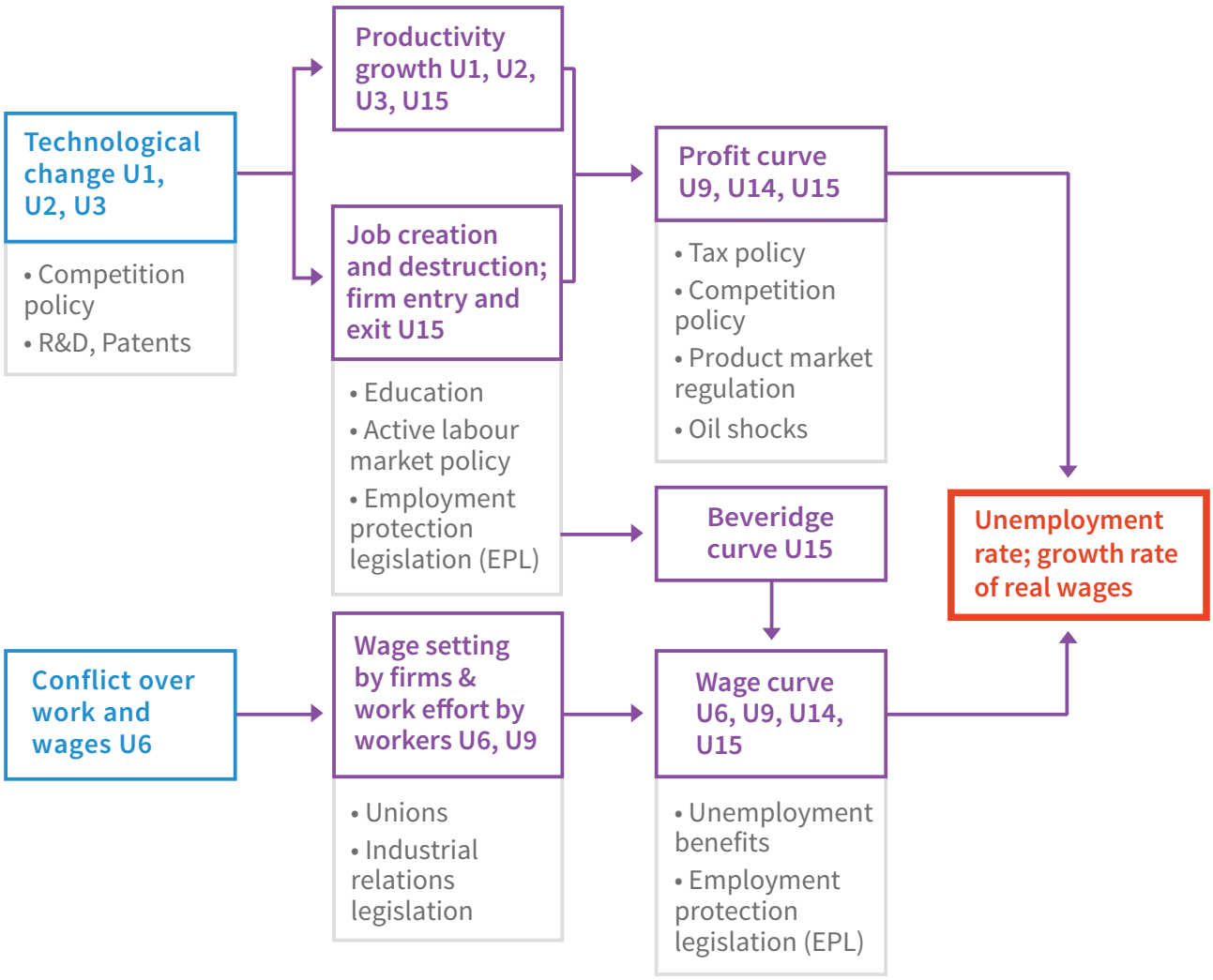


Figure 15.21 *The institutions, policies and shocks that can influence unemployment and real wages.*

Unemployment is a market failure. The pie to be divided between workers and their employers is not as large as it could be. It could be increased if more people were employed. The unemployed who would prefer to be at work are involuntarily unemployed. That's what it means when we say that employed workers are receiving a rent, they are better off than their reservation position (which is being out of work like the currently unemployed).

But firms will not hire the unemployed because, at higher levels of employment, the wages required to induce workers to work hard would drive the firms' profit rates below what is necessary to sustain investment in the firms. So the pie is smaller than it might be because of a conflict of interest over how it should be divided.

The main difference between the stars and the laggards is that, in high-performing economies, institutions and policies work so that the incentives of the main actors are to increase the size of the pie, rather than wasting resources fighting over the size of the slice.

Key points in Unit 15

Increase in output per hour of labour

Made possible by increased capital intensity of production, and advances in technologies and forms of business organisation.

Higher wages

These processes together permitted wages to rise without the profit rate falling.

Creative destruction

The introduction of labour-saving technology both destroys and creates jobs.

National economies differ

They differ in how rapidly they adjust to the opportunities offered by technological change and in the level of wages and unemployment that they can sustain in the long run.

Long-run improvement

The ability to do this requires sustained investment and innovation and so depends on the position of the profit curve as well as the wage curve.

Improvements in wages and employment

One aspect of national economic performance over the long run can be measured by the rate of increase in real wages and the rate of unemployment.

Institutions matter

Over the past 50 years good national economic performance has been associated with institutions and policies that protect workers rather than jobs, and promote innovation and entry of new firms to markets rather than protecting established firms from competition.

15.10 READ MORE

Bibliography

1. Allen, Robert C. 2012. 'Technology and the Great Divergence: Global Economic Development since 1820.' *Explorations in Economic History* 49 (1): 1–16.
2. Andersen, Torben M, Bengt Holmström, Seppo Honkapohja, Sixten Korkman, Hans Tson Söderström, and Juhana Vartiainen. 2007. *The Nordic Model: Embracing Globalization and Sharing Risks*. Helsinki: Taloustierto Oy.
3. Barciela López, Carlos, Albert Carreras, and Xavier Tafunell. 2005.Estadísticas *Históricas de España: Siglos XIX-XX*. Bilbao: Fundación BBVA.
4. Bentolila, Samuel, Tito Boeri, and Pierre Cahuc. 2010. 'Ending the Scourge of Dual Markets in Europe.' *VoxEU.org*. July 12.
5. Blanchard, Olivier. 2004. 'The Economic Future of Europe.' *Journal of Economic Perspectives* 18 (4): 3–26.
6. Blanchard, Olivier, and Justin Wolfers. 2000. 'The Role of Shocks and Institutions in the Rise of European Unemployment: The Aggregate Evidence.' *The Economic Journal* 110 (462): 1–33.
7. Blanchflower, David G, and Andrew J Oswald. 1995. 'An Introduction to the Wage Curve.' *Journal of Economic Perspectives* 9 (3): 153–67.
8. Burda, Michael, and Jennifer Hunt. 2011. 'The German Labour-Market Miracle.' *VoxEU.org*. November 2.
9. Davis, Steven J, R. Jason Faberman, and John C Haltiwanger. 2012. 'Recruiting Intensity during and after the Great Recession: National and Industry Evidence.' *American Economic Review* 102 (3): 584–88.
10. Davis, Steven J, R. Jason Faberman, and John Haltiwanger. 2012. 'Labor Market Flows in the Cross Section and over Time.' *Journal of Monetary Economics* 59 (1): 1–18.
11. Gordon, Robert J. 2004. 'Why Was Europe Left at the Station When America's Productivity Locomotive Departed?' *NBER Working Papers 10661*.
12. Habakkuk, John. 1967. *American and British Technology in the Nineteenth Century: The Search for Labour Saving Inventions*. United Kingdom: Cambridge University Press.
13. Haltiwanger, John, Stefano Scarpetta, and Helena Schweiger. 2014. 'Cross Country Differences in Job Reallocation: The Role of Industry, Firm Size and Regulations.' *Labour Economics* 26 (January): 11–25.
14. Hobsbawm, Eric, and George Rudé. 1969. *Captain Swing*. London: Lawrence and Wishart.
15. Howell, David R, Dean Baker, Andrew Glyn, and John Schmitt. 2007. 'Are Protective Labor Market Institutions at the Root of Unemployment? A Critical Review of the Evidence.' *Capitalism and Society* 2 (1).

16. Mill, John Stuart. (1848) 1909. *Principles of Political Economy: With Some of Their Applications to Social Philosophy*. London: Longmans Green and Co.
17. Nelson, Richard R, and Gavin Wright. 1992. 'The Rise and Fall of American Technological Leadership: The Postwar Era in Historical Perspective.' *Journal of Economic Literature* 30 (4). American Economic Association: 1931–64.
18. Nickell, Stephen, and Jan van Ours. 2000. 'The Netherlands and the United Kingdom: A European Unemployment Miracle?' *Economic Policy* 15 (30): 136–80.
19. OECD. 2015. 'OECD Statistics.'
20. Raff, Daniel. 1988. 'Wage Determination Theory and the Five-Dollar Day at Ford.' *The Journal of Economic History* 48 (02): 387–99.
21. Rifkin, Jeremy. 1996. *The End of Work: The Decline of the Global Labor Force and the Dawn of the Post-Market Era*. New York, NY: G.P. Putnam's Sons.
22. Russell, Bertrand. 1935. *In Praise of Idleness and Other Essays*. London: George Routledge.
23. Singer, Natasha. 2014. 'In the Sharing Economy, Workers Find Both Freedom and Uncertainty.' *The New York Times*, August 22.
24. Sterk, Vincent. 2015. 'Home Equity, Mobility, and Macroeconomic Fluctuations.' *Journal of Monetary Economics* 74 (September): 16–32.
25. The Conference Board. 2013. 'International Comparisons of Hourly Compensation Costs in Manufacturing.'
26. US Bureau of Labor Statistics. 2004. 'International Labor Comparisons (ILC).' October 14.
27. Visser, Jelle. 2013. 'ICTWSS: Database on Institutional Characteristics of Trade Unions, Wage Setting, State Intervention and Social Pacts in 51 Countries between 1960 and 2014.' *Amsterdam Institute for Advanced Labour Studies (AIAS)*.
28. Wooldridge, Adrian. 2013. 'Northern Lights.' *The Economist*. February 2.

16

THE NATION IN THE WORLD ECONOMY

HOW THE INTEGRATION OF NATIONAL ECONOMIES INTO A GLOBAL SYSTEM OF TRADE AND INVESTMENT PROVIDES OPPORTUNITIES FOR MUTUAL GAINS—AND CONFLICTS OVER THE DISTRIBUTION OF THE GAINS

- Globalisation includes the integration of markets in goods and services and investment, while labour markets remain largely national
- Globalisation of goods and services has led to the prices of goods being more similar in different countries, but wages show no corresponding tendency towards convergence
- Nations tend to specialise in the production of the goods and services in which they are relatively low-cost producers, for example because of an abundance of resources or skills
- The goods in which a nation is a low-cost producer change over time, in response to public policies
- This specialisation allows for mutual gains for the people of trading countries
- The distribution of these gains among countries may favour those with superior bargaining power
- Within nations some sectors of the economy, and the owners of some factors of production, benefit while others lose—at least in the short run
- In the long run, fully exploiting the mutual gains made possible by international specialisation and investment, and distributing those gains fairly, will depend on the institutions and policies that nations adopt

Institute for New Economic Thinking
UCL SciencesPo Azim Premji University
FRIENDS PROVIDENT Foundation

In December 1899 the steamship *Manila* docked in Genoa, Italy, and offloaded her cargo of grain grown in India. The Suez Canal had opened 30 years earlier, slashing the cost of moving agricultural commodities from south Asia to European markets. Italian bakers and shoppers were delighted at the low prices. Italian farmers were not. After a couple of months in Genoa, the *Manila* headed west. She carried 69 people in steerage (the cheapest possible passage) who were abandoning their homeland in search of a livelihood in the US.

The low prices were made possible by a revolution in transportation and farming technology. Like the opening of the Suez Canal, the expansion of the railway system to the grain fields of North America, the Russian plain and the farmlands of northern India, and the development of steam powered ships like the *Manila* cut the cost of transporting grain to distant markets. Across the vast plains of the American midwest, new strains of wheat, newly developed reapers and sowers, and improved drainage technologies had created a hi-tech, capital-intensive form of farming that was as productive as anywhere in the world.

Across Europe parliaments and state bodies struggled to adjust to the grain price shock. In France and Germany farmers and their advocates prevailed. Despite the benefits of lower grain prices to families, and despite the protests of workers who consumed the grain, governments imposed *tariffs* to protect the incomes of farmers.

Denmark, among other countries, responded differently. Instead of protecting grain farmers from cheap imports, the government helped them to start dairy farming instead. Using cheap imported grain as an input, drawing on a tradition of cooperation, farmers responded to the incentives to produce milk, cheese and other commodities that could not be transported cheaply over long distances. In turn, cheaper grain meant families could increase spending on these dairy products.

In Italy the children of some farmers took up jobs in the booming textile industry, which was exporting to the rest of the world. Many bankrupted farmers made the trip to the US. They slept on the decks of empty freighters that were returning to the US to pick up grain for Europe. About 750,000 Europeans made this voyage each year during the decade after the *Manila*'s visit to Genoa. Some of their grandchildren would end up as American farmers, growing grain in Kansas.

There were big winners and big losers from the grain price shock. Many of the changes made economic sense: for example, the world's grain was now grown increasingly in places where it could be produced most efficiently. But tariffs designed to protect the farmers of Germany and France held back this reallocation, preventing owners and workers in other sectors of the economy from enjoying lower grain prices. This continues to this day: rich countries still commonly protect their agricultural sectors through subsidies.

The battle line was not between rich and poor, or landlords and tenants, or employers and employees. The conflict was between the producers of different commodities: those involved in manufacturing welcomed the expansion of trade with the US, while those farming grain did not.

The word commonly used to describe our increasingly interconnected world is *globalisation*. The term refers not only to the trade in grain and migration across national borders illustrated by the Manila, but also to non-economic aspects of international integration: the International Criminal Court, the flow of ideas across borders, or our increasingly similar taste in music that we mentioned in Unit 1.

In Unit 6 we looked at firms like Apple that offshore production to other parts of the world where costs are lower. This *offshoring* is an important dimension of globalisation, and it can involve outsourcing production to other companies, or it can take place within the boundaries of a multinational company. For example, Figure 16.1 shows that the Ford Motor Company operates offices or plants in 22 countries outside the US. The company started offshoring a year after it was founded (in Canada in 1904) and started manufacturing in many other countries soon afterwards, for example in Australia (1908) and even the Soviet Union (1930). This "American" company has almost 190,000 employees. More than 142,000 of them are located outside the US.

GLOBALISATION

A process by which the economies of the world become increasingly integrated by the freer flow across national boundaries of:

- Goods
- Investment
- Finance
- Labour (to a lesser extent)

The term is sometimes applied more broadly to include ideas, culture, and even the spread of epidemic diseases.

In the case of a multinational company, owners, managers and employees in many countries have become part of the same unified, transnational structure. This is because the costs of doing business within the company are lower than the costs of doing business between companies. But, as we saw with the cotton market in Unit 9, globalisation not only involves the integration of firms in different countries; it involves the integration of markets themselves, bringing sellers and buyers in different countries closer together.

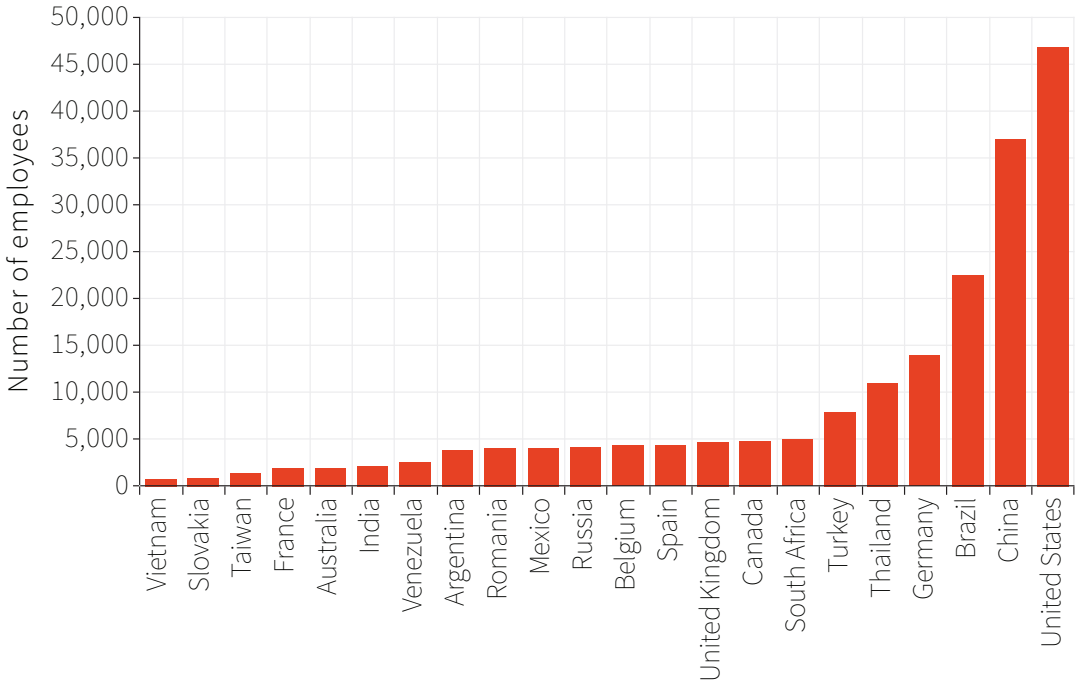


Figure 16.1 *Ford employees across the world in 2014.*

Source: *Ford Motor Company. Note: Ford has factories in Japan but does not provide employee numbers for them, so they have been omitted.*

You have already learned the basic concepts that you need to understand the global economy:

- *Exchange* involves the possibility of mutual gains and also conflicts over how these gains will be distributed.
- *Competition* typically leads to an outcome in which all concerned—whether they are consumers, employees, owners of firms or others—are doing the best they can under the circumstances.
- The resulting *Nash equilibria* may not be Pareto efficient and in the eyes of many may seem unfair; government policies may improve the result either from the standpoint of efficiency or fairness, but this outcome is not at all guaranteed.

When a globalised economy means that goods, services, people, and financial assets often cross national boundaries, this introduces other factors. Governments have additional powers and policies that include:

- *Imposition of tariffs*: Effectively they discriminate against goods produced elsewhere.
- *Immigration policies:* Governments regulate the movement of people between nations in a way that would not be possible (or acceptable) within most nations.
- *Capital controls:* Limits on the ability of individuals or firms to transfer financial assets among countries.

- *Monetary policies:* They affect the exchange rate, and so alter the relative prices of imported and exported goods.

While national boundaries give governments additional policy tools, they also limit the reach of governments. Within a nation, governments generally succeed in protecting private property rights where these exist, and in enforcing contracts. Because there is no world government (and rather weak international agencies), enforcing contracts and protecting property rights globally is sometimes impossible.

This raises controversial questions about the fairness of the distribution of mutual gains from exchange. The conflicting interests sometimes coincide with national differences between poorer and richer economies: it is tempting, though often inaccurate as we will see, to consider these conflicts as "us" at home versus "them" abroad.

In this unit we will consider three markets that became more integrated with globalisation: international markets for goods and services (trade); international labour markets (migration); and international capital markets (international capital flows, which are flows of savings and investment).

16.1 GLOBALISATION AND DEGLOBALISATION IN THE LONG RUN

The trade in goods, sometimes called *merchandise trade*, concerns tangible products that are physically shipped across borders by road, rail, water or air. Trade of this sort has been happening for millennia, although the nature of the goods traded, and the distances over which they have been shipped, have changed dramatically. Trade in services is a more recent phenomenon, although it has also been occurring for centuries. Services that are commonly traded across borders are tourism, financial services and legal advice. Many traded services make merchandise trade easier or cheaper—shipping services, insurance and financial services for example.

The UK became the leading provider of these services during the 19th century, when it was the most advanced industrial economy, the major naval and imperial power, and the most important trading nation. Nowadays, countries also export educational services (for example, people come from all over the world to study in US or European universities), consulting services and medical services. India has become a major exporter of software-related services: for example, this project was developed in Bangalore. We will study these service exports together with merchandise trade, since the same principles can help us to understand them.

How can we measure the extent of globalisation in goods and services? One approach would simply be to measure the amount of trade in a country or region, or the world as a whole, over time. If it increased, we conclude that the country, region or world was becoming more globalised. It is common to measure trends in the share of imports, or exports, or total trade (imports plus exports) in GDP as an indicator of globalisation, so as to take account of the growth of GDP as well as trade.

Figure 16.2 shows world merchandise exports, expressed as a share of world GDP, between 1820 and 2011. The share rose by a factor of eight, from 1% to 8%, between 1820 and 1913. In 1950, the share was lower (5.5%) but recovered rapidly during the prosperous post-war period. It reached 10.5% in 1973, 17% in 1998, and 26% in 2011. In the long run the trend has clearly been upwards, with a sharp acceleration from the 1990s onwards. However, this trend was interrupted between 1914 and 1945, which included two world wars and the Great Depression.

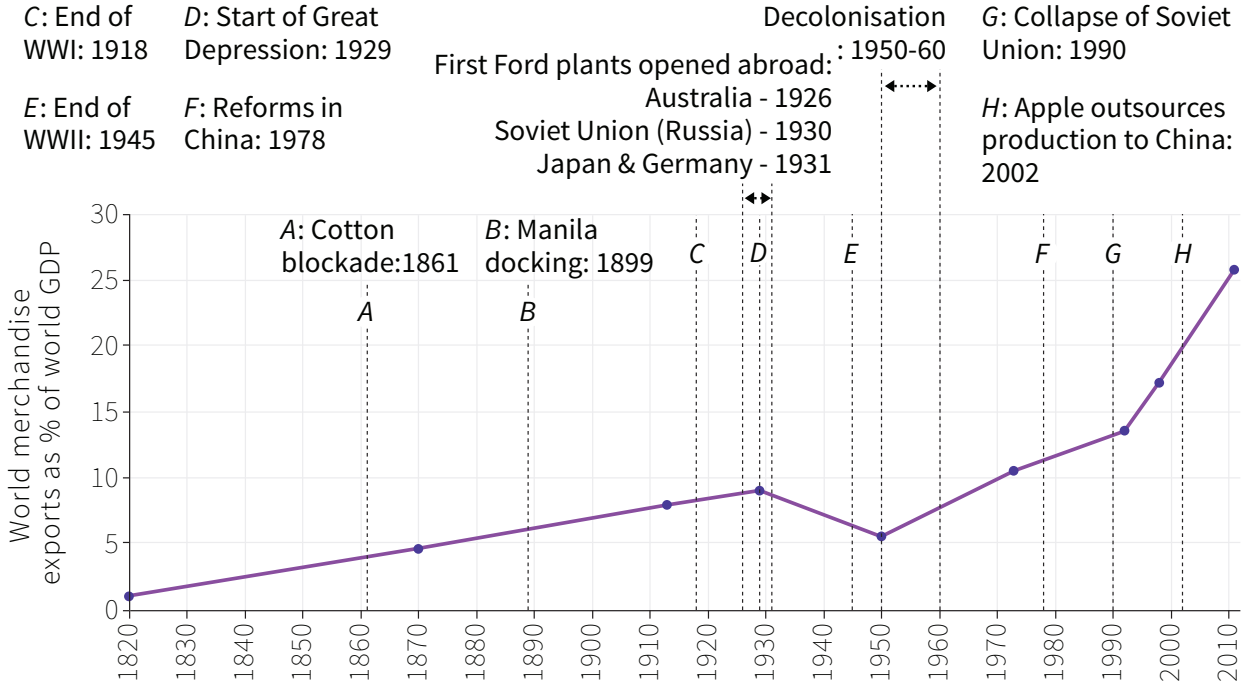
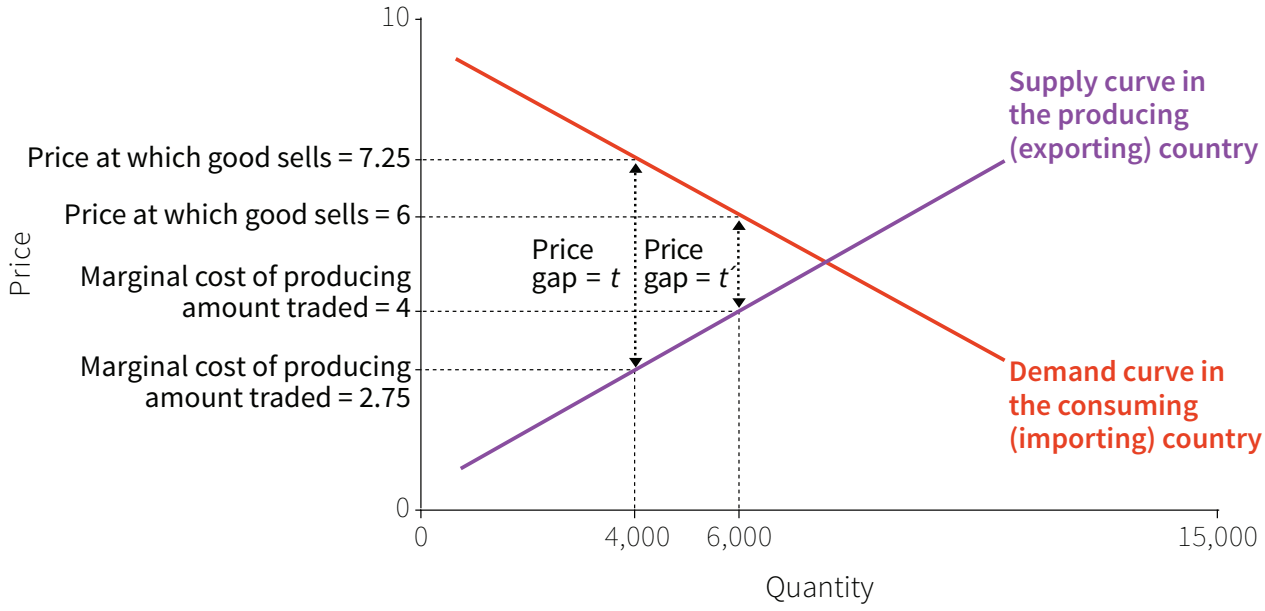


Figure 16.2 *World merchandise exports as a share of world GDP (1820-2011).*

Source: *Appendix I in Maddison, Angus. 1995. Monitoring the World Economy, 1820-1992. Washington, DC: Development Centre of the Organisation for Economic Co-operation and Development. Table F-5 in Maddison, Angus. 2001. The World Economy: A Millennial Perspective (Development Centre Studies). Paris: Organisation for Economic Co-operation and Development. World Trade Organization. 2013. World Trade Report. Geneva: WTO.International Monetary Fund. 2014. World Economic Outlook April: Recovery Strengthens, Remains Uneven. Washington, DC: IMF.*

Measuring the reduction in *trade costs* among countries is a second method. When the costs of trading between countries fall then, in economic terms, the world has shrunk; it is as if countries were closer. In Unit 8 you learned about Alfred Marshall. We saw that the *Law of one price* holds in markets with many potential buyers and sellers, where all goods are identical and where buyers and sellers are aware of all trading opportunities. But this assumes that it is costless to take advantage of

those trading opportunities. If, on the other hand, trading between markets in two countries is costly because of transport costs, trade barriers or other factors, then there is no reason to suppose that prices will be the same in both.



Let us assume that the cost of shipping a unit of the good is 4.5. We will show that 4,000 units will be produced.

Why 4,000?

Because at that quantity, the difference between the supply curve and the demand curve is equal to the trade cost, 4.5. The marginal cost in the producing country will be 2.75, while the importing country is willing to pay 7.25 per unit.

The effect of globalisation

If we think of globalisation as a process, then a world that is becoming more globalised is one in which trade costs are falling. In the figure, this is represented by a decline in trade costs from t to t'.

The price gap declines

As can be seen, falling trade costs imply a decline in the price gap between the import price and the export price and an increase in the quantity traded from 4,000 to 6,000.

Figure 16.3 *The market for a traded good: Price gaps reflect trade costs.*

Consider the market for a good that is produced in (and exported from) one country and consumed in (and imported into) another. To keep the analysis straightforward, imagine that these are the two only countries in the world; that none of the good

is consumed in the producing country; and that none of the good is produced in the consuming country. This means that everything that is produced is traded. The purple line in Figure 16.3 represents the supply curve in the producing (exporting) country: it is an upward-sloping function of the price in that country. The red line represents the demand curve in the consuming (importing) country: it is a downward-sloping function of the price in that country.

Let t be the cost of shipping a unit of the good from the exporting to the importing country, including all transportation costs, trade taxes and so on. It is a measure of the *price gap* between the price the good in the exporting country and the price of the good in the importing country. If the market is competitive, then the total cost of obtaining a unit of the good in the importing country will be the cost of buying it in the exporting country, plus the trade cost t. Follow the slideline to see how changes in trade costs are reflected in price gaps.

PRICE GAP

A difference in the price of a good in the exporting country and the importing country. The price gap includes transportation costs and trade taxes. When global markets are in competitive equilibrium, price gaps will be entirely due to trade costs.

- Globalisation benefits both exporting producers and importing consumers.
- It does so by bringing them closer together, and it leads to an increase in both the supply of exports and the demand for imports.

The concept of *arbitrage* explains why the price gap should tend to equal the sum of all trade costs. By buying at a low price in export markets and selling at a higher price in import markets, traders can make a profit as long as the price gap is higher than the total costs of trade. When traders engage in arbitrage in this fashion, they lower the supply of the good in the export market, driving up its price; and they increase the supply of the good in the import market, lowering its price, and causing the price gap to decline. This should continue until price gaps have been driven down to the trade cost, and further arbitrage is unprofitable. A high price gap reflects a world in which trade is expensive and globalisation is limited. A low price gap, on the other hand, reflects a much more globalised world in which trade is cheap.

This means we can learn about globalisation from data on prices:

- *Globalisation should lead to falling import prices:* But if we observe falling import prices this does not necessarily mean that globalisation is occurring. The demand for the good in question may simply have declined (or the supply may have increased).
- *Globalisation should also lead to rising export prices:* But rising export prices do not necessarily imply globalisation. Demand for the good in question may simply be rising (or the supply may have declined).

- *Declining price gaps between importing and exporting countries are a much surer sign of globalisation:* Especially if they are accompanied by rising trade volumes.

For example, Figure 16.4 shows unmistakable evidence of declining transatlantic trade costs during the 19th century. The wheat price gap between the UK and the US (expressed as a percentage) fluctuated wildly before 1840 or so, around a roughly constant trend. It then started to decline at about the same time that shipping costs started to fall, a result of the introduction of steamships on long-distance routes. The price gap had almost vanished by 1914. At the same time, the volume of wheat shipped across the Atlantic rose dramatically.

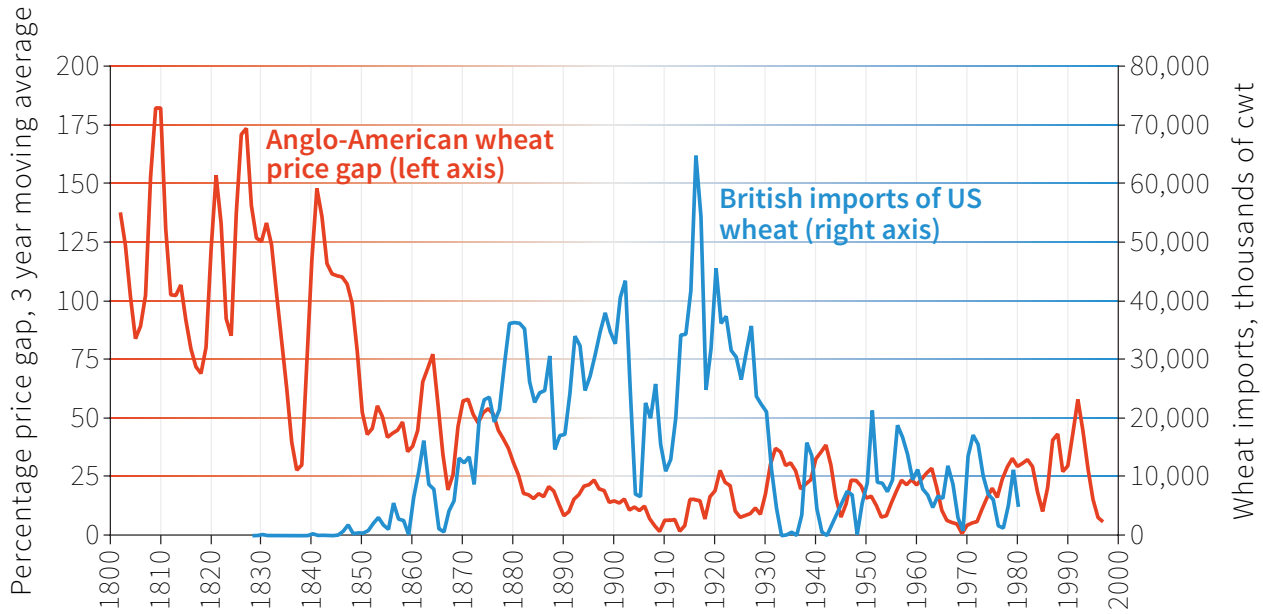


Figure 16.4 *The Anglo-American wheat trade (1800-2000).*

Source: *Figure 3 in O'Rourke, Kevin H., and Jeffrey G. Williamson. 2005. 'From Malthus to Ohlin: Trade, Industrialisation and Distribution since 1500.' Journal of Economic Growth 10 (1): 5–34.*

The transatlantic trade in wheat is not an isolated example. International price gaps fell sharply on many routes and for many commodities between 1815 and 1914, the first epoch of modern globalisation.

Figure 16.5 gives "American-Anglo" price gaps (the reverse of Figure 16.4) for a variety of commodities between 1870 and 1913. For agricultural commodities such as wheat and animal products, British prices were higher than American ones, so the price gaps are the percentage by which the British price exceeded the American price. In the case of industrial commodities such as cotton textiles or iron bars, American prices were higher than British ones, so the price gaps quoted are the percentage by which prices in Boston or Philadelphia exceeded prices in Manchester or London. In nearly all cases (sugar is the outstanding exception) price gaps fell, indicating that transatlantic commodity markets were becoming better integrated. Much like the dramatic reduction in grain prices in Genoa after the opening of the

Suez Canal, which we discussed in the introduction to this unit, price gaps between the US and the UK reduced over time because of a revolution in transportation and improvements in farming and production technology.

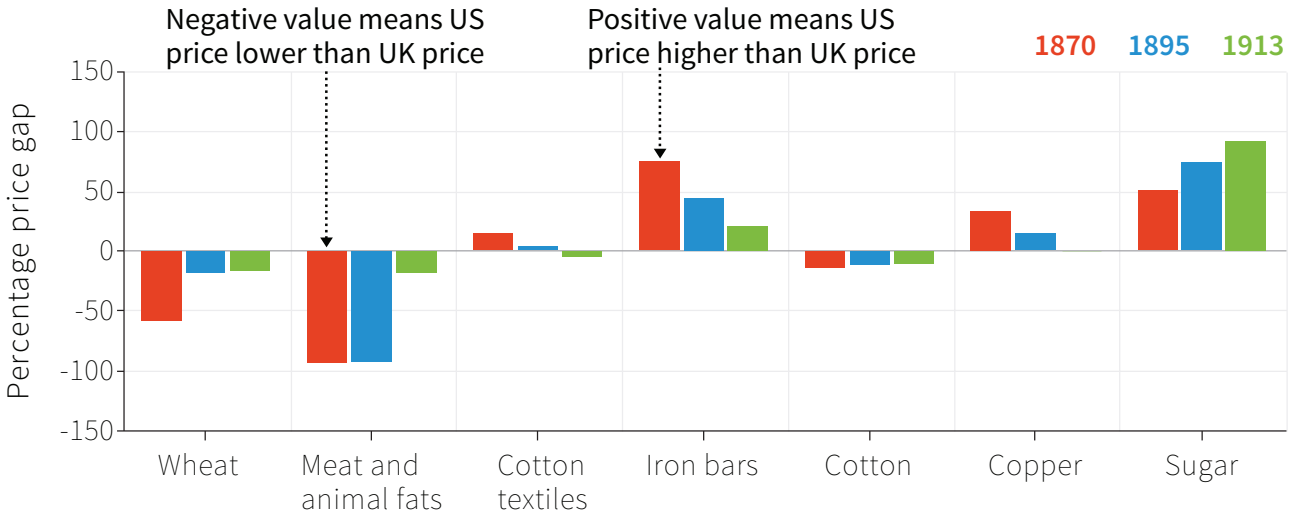


Figure 16.5 *Commodity price gaps between the US and UK (1870-1913).*

Source: *Table 2 in O'Rourke, Kevin, and Jeffrey G. Williamson. 1994. 'Late Nineteenth-Century Anglo-American Factor-Price Convergence: Were Heckscher and Ohlin Right?' The Journal of Economic History 54 (04): 892–916.*

Nor was price convergence limited to the north Atlantic. Between 1873 and 1913, the Liverpool-Bombay cotton price gap fell from 57% to 20%; the London-Calcutta jute price gap fell from 35% to 4%; and the London-Rangoon rice price gap fell from 93% to 26%. Between 1846 and 1855, and between 1871 and 1879, during which period Japan was opened up to trade with the rest of the world, the Japan-Hamburg nail price gap fell from 400% to 32%, and the refined sugar price gap fell from 271% to 39%.

DISCUSS 16.1: PRICE GAPS THAT DIDN'T FALL

Figure 16.5 shows the price gap of different commodities between the US and UK over time. Can you think of a reason why price gaps for meat and animal fats such as butter did not start to fall until 1895?

Railways were probably even more important than steamships in integrating global commodity markets—without them it would have been prohibitively expensive to ship grain and other goods from the interior of continents to coastal seaports. Where price gaps fell less sharply during the late 19th century, this was often because of tariffs—taxes on imports—which were rising in several countries for reasons that we will discuss later, and which counteracted the effects of declining transport costs.

Figure 16.4 suggests that transatlantic shipments of wheat fell after 1914, and that price gaps rose, suggesting a rise in trade costs and therefore deglobalisation. International price gaps rose during the interwar period for many agricultural commodities, because governments raised tariffs in response to unemployment and economic insecurity. When a country undertakes *protectionist policies*, its government is taking steps to limit trade—in particular to reduce the amount of imports coming into the economy. This is often done to protect domestic industries against foreign competition (hence protectionism), but consumers have to pay more for imports. Protectionist measures include taxes to raise the domestic price of imports (a tariff) and quantitative restrictions on imports (a quota).

Figure 16.4 suggests that the post-1945 period was one of "reglobalisation": slow at first, but then accelerating, especially after 1990. Agricultural markets were largely protected for much of the period, and there is no reason to suppose that international price gaps for agricultural commodities fell sharply. The markets for industrial goods and components, on the other hand, were liberalised, and several studies have found evidence of declining international price gaps in the late 20th century.

Economists have measured trade costs indirectly, by looking at trade between pairs of countries. This shows long-term changes in impediments to trade, and can separate the contribution of distance between the countries from national policies of those countries. If trade between Germany and France, for example, increased from one year to the next, but it did not increase between these two countries and their other trading partners at the same time, we can interpret this as an indirect measure of declining trade costs for this pair of countries.

If we sum total trade costs each year for all major economies, we have an indicator of the process of globalisation. Figures 16.6 and 16.7 do this for the period 1870 to 2000.

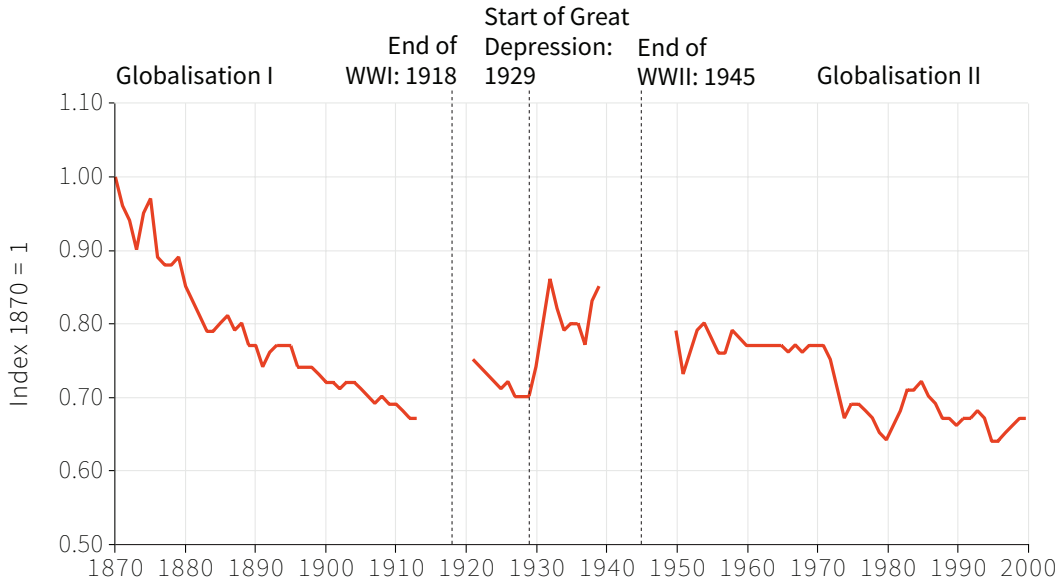


Figure 16.6 *Impediments to trade (1870-2000).*

Source: *Jacks, David S., Christopher M. Meissner, and Dennis Novy. 2011. 'Trade Booms, Trade Busts, and Trade Costs.' Journal of International Economics 83 (2): 185–201. Note: Presented as an index, with 1870 = 1.*

Trade costs declined substantially from 1870 to 1913, reflecting declining transport costs and reductions in tariffs. Trade costs then rose in the interwar period because of rising tariffs. This was particularly the case following the onset of the Great Depression in 1929: countries attempted to solve unemployment problems by discouraging imports.

From 1970 trade costs started to fall worldwide again as countries began to re-liberalise trade, and transport technologies improved. Tariffs tend to be higher in low-income countries than in rich countries, but recently most countries have reduced their tariffs.

The price evidence therefore suggests interrupted commodity market integration over the past 150 years: 19th century integration was briefly reversed, followed by reintegration after the second world war. We call these two periods of integration *globalisation I and globalisation II*.

GLOBALISATION I AND II

Two separate periods of increasing global economic integration:

- Globalisation I extended from before 1870 until the outbreak of the first world war in 1914
- Globalisation II extended from the end of the second world war into the 21st century

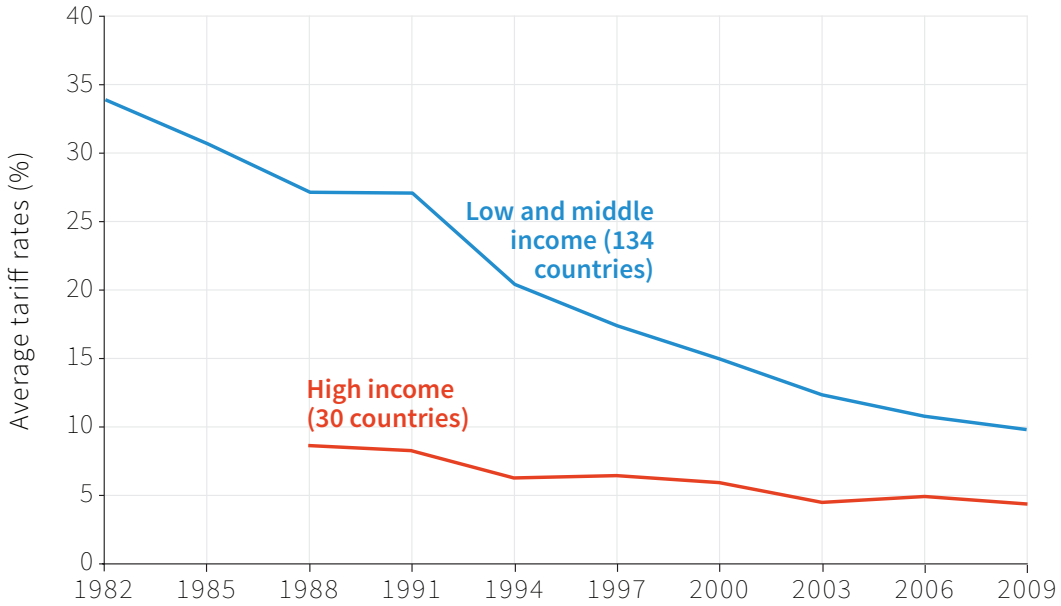


Figure 16.7 *Average tariff rates, percent (1981-2010).*

Source: *The World Bank. 2011. 'Data on Trade and Import Barriers.' Note: 3-year centred moving average.*

DISCUSS 16.2: LEARNING MORE ABOUT TARIFFS

Download the World Bank data set *Trends in average MFN applied tariff rates in developing and industrial countries, 1981-2010* used to produce Figure 16.7.

1. Choose one country from each income category (code 1 to 4) and plot the evolution of tariffs in these four countries. Use your plots to describe how tariffs in your sample have changed over time.
2. The empirical evidence suggests that on average, tariffs tend to be higher in lower income countries than in rich countries, but that most countries have reduced tariffs substantially in recent decades. Discuss why your plots support (or don't support) this claim for your chosen countries.

(As a starting point you may want to consider your chosen countries' membership of global trade agreements such as GATT/WTO or the EU, and also whether the country has followed the structural adjustment programmes of the IMF.)

16.2 GLOBALISATION AND INVESTMENT

In international capital markets there is a similar pattern of 19th century globalisation, followed by a brief episode of interwar deglobalisation, and late 20th century reglobalisation.

If countries existed in isolation they would have to finance their investment needs using their own savings. If this were the case, they could not spend more than they earned in a year, and all their income would have to be spent domestically: domestic expenditure would have to equal domestic income. Individuals, financial institutions, companies and governments in one country, however, can lend to individuals, financial and non-financial companies and governments in another. To keep the language simple, let's talk about *countries* lending to or borrowing from other countries, bearing in mind the fact that these countries are made up of many individuals, companies and institutions. A country can spend more than it earns by borrowing from abroad. Similarly a country can decide not to use its savings to finance domestic investment, and instead lend it abroad and earn a return on these foreign loans. In this case its savings will exceed domestic investment, or (equivalently) its income will be higher than expenditure.

We use the *balance of payments* accounts to track lending and borrowing abroad. We first need to explain how lending and borrowing abroad is related to international trade in goods and services. This is because imports represent payments from the domestic economy to the rest of the world, while exports represent payments from the rest of the world to the domestic economy. The balance of payments records the sources and uses of foreign exchange: if the records of transactions were complete, the balance would sum to zero because the source and use of each dollar crossing an international border could be accounted for. (An entry called *errors and omissions* is added to the balance of payments accounts to make it sum to zero.)

To see how the balance of payments accounts work, think first about an economy where the only international payments are due to trade. If the home country imports more than it exports, then its residents are making more international payments than they are receiving. For example, a country buying a higher value of imports from the US than it receives from selling its exports to the US needs to get hold of the dollars—by borrowing from the US or the rest of the world—to cover the difference.

Conversely, if the home country is exporting more than it is importing, then its nationals must be lending to their trading partners so they can pay for the exports. These loans are a use of foreign exchange for the home country, and a source of foreign exchange for its trade partners.

Thus a trade deficit will imply that the country is borrowing, while a trade surplus implies it is lending (which we saw in Unit 11 is equivalent to saving).

There are other reasons that people in one country make payments to people in another. The most important is the purchase of assets in another country. If a US company purchases shares in a company in China, it is making a payment for a Chinese asset. This implies a payment from the US to China. This is a use of foreign exchange called *portfolio investment*. Similarly, if a US company purchases a factory in China, this is a use of foreign exchange called *foreign direct investment* (FDI).

In subsequent years, however, the US company will receive dividends from its portfolio investment or profits from its direct investment, which will be sent back, or repatriated, to the US company. These repatriated profits are payments from China to the US. They are recorded in the US balance of payments as a source of foreign exchange.

Other important international payments include money sent home by migrant workers to their families (called *remittances*) and official aid flows, mostly from the governments of rich to poor countries.

All of these international payments are tracked in the balance of payments accounts, and the net value of these payments is called the *current account* or CA—so the CA is the sum of all payments made to a country minus all payments made by the country. A country might have a trade deficit—that is, import more than it is exporting—but still have a current account surplus if it is receiving more than enough income from

its foreign investments, remittances or foreign aid to pay the difference. In this case it will not need to borrow. For simplicity we ignore remittances and international aid and assume that the current account is equal to exports, X, minus imports, M, plus net earnings from assets owned abroad.

CURRENT ACCOUNT (CA)

The sum of all payments made to a country minus all payments made by the country.

$$current\ account \equiv exports - imports + net\ earnings\ from\ assets\ abroad$$
$$CA \equiv X - M + NINV$$

Since the current account includes all international payments, it also tells us directly whether a country is borrowing or lending:

- *CA deficit*: This means the country is borrowing—it has to do so to cover the excess payments it is making to the rest of the world.
- *CA surplus*: This means the country is lending (saving) to allow other countries to send it the excess payments.

The borrowing and lending tracked by the current account are known as *net capital flows*. In this context *capital* means money that is being lent and borrowed, rather than capital goods. A country that is borrowing (has a CA deficit) is receiving net capital flows—it is being lent cash in order to cover its CA deficit. This cash will have to be paid back in future, so capital inflows also represent rising foreign debt for the country. If the borrowed money is used for productive investments though, the investment can generate the income used to repay the debt. Thus, when a country wants to invest more than can be paid out of its own savings, borrowing from abroad can be used to finance the extra investment.

Historically, increased trade has tended to lead to larger CA imbalances. That is, when countries trade more, they also tend to borrow and lend more. The measure shown in Figure 16.8 is the sum of the absolute values of the current account balances of 15 countries from 1870 to 2014. We add up the absolute value of the current accounts to capture both borrowing and lending across countries.

The volume of capital flows in Figure 16.8 reflects a pattern of interrupted globalisation. During the late 19th century there were huge capital flows from northwest Europe where there were current account surpluses—the UK especially, but also France and Germany—which financed investment in railways and infrastructure in countries such as Argentina, Australia, Canada, and the US. These were all countries with abundant and underexploited natural resources, especially land, but railways had to be extended into the interior in order to exploit these resources, and the land had to be settled with immigrants. The investments made

a healthy return, since they increased the productive capacity of the borrowing countries, which were able to pay back the loans with interest based on the increased incomes that they enjoyed as a result. (In Europe the countries that succeeded in attracting foreign investment during this period, such as Russia and Sweden, were also relatively resource-abundant.)

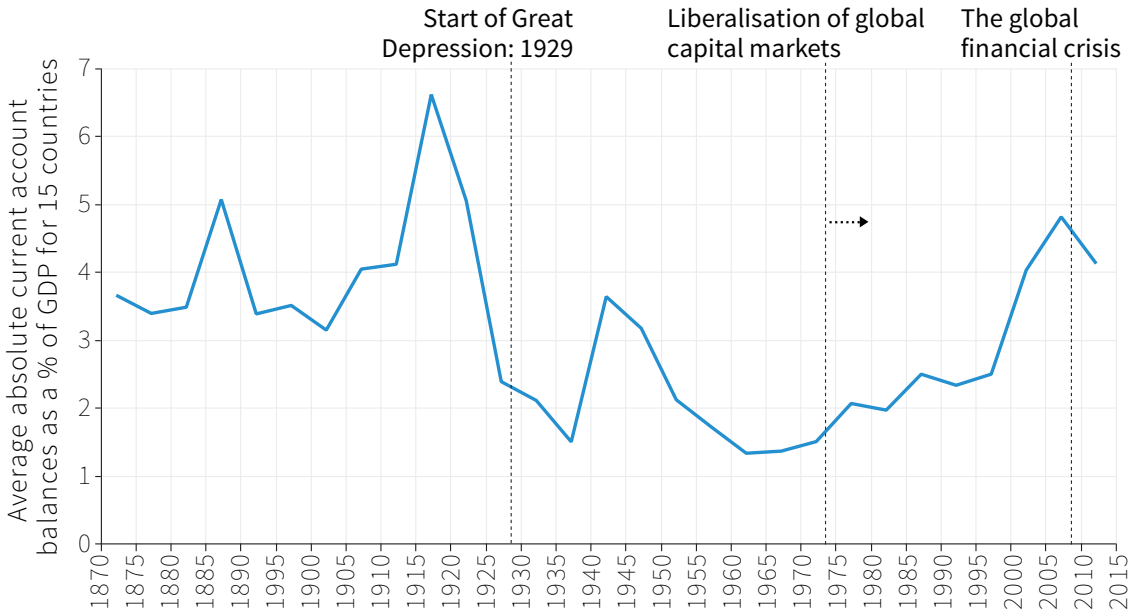


Figure 16.8 *International capital flows (1870-2014).*

Source: *Figure 2.2 from Obstfeld, Maurice, and Alan M. Taylor. 2005. Global Capital Markets: Integration, Crisis, and Growth (Japan-US Center UFJ Bank Monographs on International Financial Markets). Cambridge: Cambridge University Press. International Monetary Fund. 2014. 'World Economic Outlook Database: October 2014.' Note: The data shown in the figure is the average absolute current account balance (as a percentage of GDP) for 15 countries in five-year blocks (from 1870-74 through to 2010-14). The countries in the sample are Argentina, Australia, Canada, Denmark, Finland, France, Germany, Italy, Japan, Netherlands, Norway, Spain, Sweden, UK, US. Data for 2014 is an IMF estimate.*

In the interwar period these capital flows fell sharply—especially after the beginning of the Great Depression in 1929, which had led many countries to impose strict limits on the movement of capital across frontiers. These limits on capital flows meant that countries had to keep their current account deficits and surpluses relatively low—they prevented the capital inflows that would be required to finance large current account deficits. Unlike international trade, which resumed growth soon after the end of the second world war, capital controls persisted for longer and only started to be relaxed in the 1970s and 1980s. Since then capital flows have increased sharply. Net capital flows as measured by current account balances, representing the transfer of savings from one country to another, have been as large over the last 20 years as they were before the first world war. Gross capital flows, representing the two-way flows of speculative capital chasing short-run arbitrage opportunities, have been much higher.

Figure 16.9 shows how international asset holdings evolved during the 20th century. The pattern is U-shaped. For the rich countries that dominated international lending the share of foreign assets divided by GDP was high in the early part of the century,

but collapsed in the 1930s. After 1945, New York took over from London as the global financial centre and the US eclipsed Britain as the dominant international asset holder.

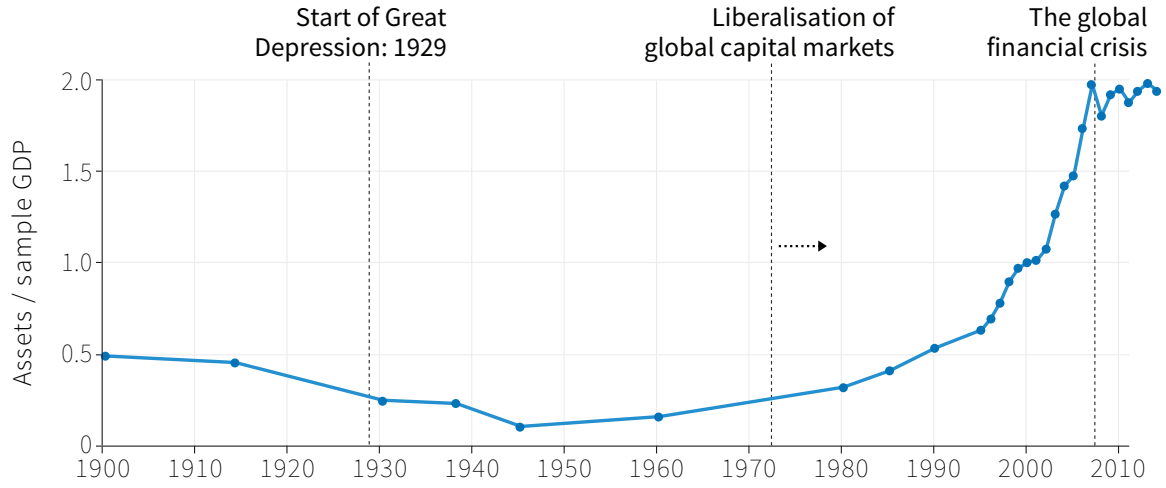


Figure 16.9 *International asset holdings (1900-2014).*

Source: *Figure 2.2 from Obstfeld, Maurice, and Alan M. Taylor. 2005. Global Capital Markets: Integration, Crisis, and Growth (Japan-US Center UFJ Bank Monographs on International Financial Markets). Cambridge: Cambridge University Press. Lane, Philip R., and Gian-Maria Milesi-Ferretti. 2007. 'Europe and Global Imbalances.' IMF Working Papers 07 (144).*

DISCUSS 16.3: INTERNATIONAL CAPITAL FLOWS AND CHINA

1. China has enjoyed a period of rapid development over recent decades. Describe how China's current account balance has evolved since the late 1990s (you can find data for 1998-2012 at FRED (search for "total current account balance for China").
2. Now compare China's current account balance from your previous answer to that of the US for the same time period (you can find the data for the US at FRED by searching for "total current account balance for United States").
3. Look at Figure 16.8 and note that international capital flows (as measured by average absolute current account balances as a proportion of GDP) in the first decades of the 21st century are similar to those of the late 19th century. By considering your answers to the two questions above, and using our description of capital flows in the late 19th century when countries like Argentina and the US were growing fast, explain whether the pattern of international capital flows in the 21st century is similar to that of the late 19th century.

To measure price gaps in international capital markets, we need prices of identical financial assets in different countries. Where researchers have been able to locate such prices, they have found that the late 19th century saw significant globalisation in capital flows.

For most of the 19th century, if a would-be arbitrageur in New York (or London) wished to act on a price gap between New York and London, the speed at which information travelled limited his opportunities (we tracked the speed at which information travelled over the past 1,000 years in Unit 1). Information about price gaps travelled on ships that crossed the Atlantic. By the time the arbitrageur learned of the price gap, the information was already several days out of date. To act on it, he had to send written instructions to his agent in the other city to buy or sell. These instructions travelled by ship too. It took a large price gap, therefore, to create speculation.

In 1866, investors in London and New York (and their agents) could, for the first time, communicate with each other on the same day. They were using the first transatlantic telegraph cable, running from Ireland to Newfoundland in Canada. Once the cable was installed, investors could act immediately when they heard of a potential arbitrage opportunity, and price gaps immediately collapsed.

In most countries, residents and companies do the majority of investment. One dimension of globalisation is the foreign direct investment (FDI) mentioned earlier by companies abroad, including subsidiaries. Unlike the use of savings to buy foreign bonds or shares in a foreign company (portfolio investment), the intention of FDI is to exercise control over the use of resources in the foreign company.

Figure 16.10 shows the destination of investments by US companies when they invested abroad between 2001 and 2012. Perhaps surprisingly, when US firms chose to produce outside the US, they predominantly went to countries in Europe; and in Europe largely to countries in which wages were higher than in the US. The Netherlands, Germany and the UK alone received more US investment than Asia and Africa combined. In this respect the location of Ford plants around the world shown in Figure 16.1 is not typical, because Ford has far more employees in China, Brazil, Thailand and South Africa combined than in Germany, UK, Canada, Belgium and France combined.

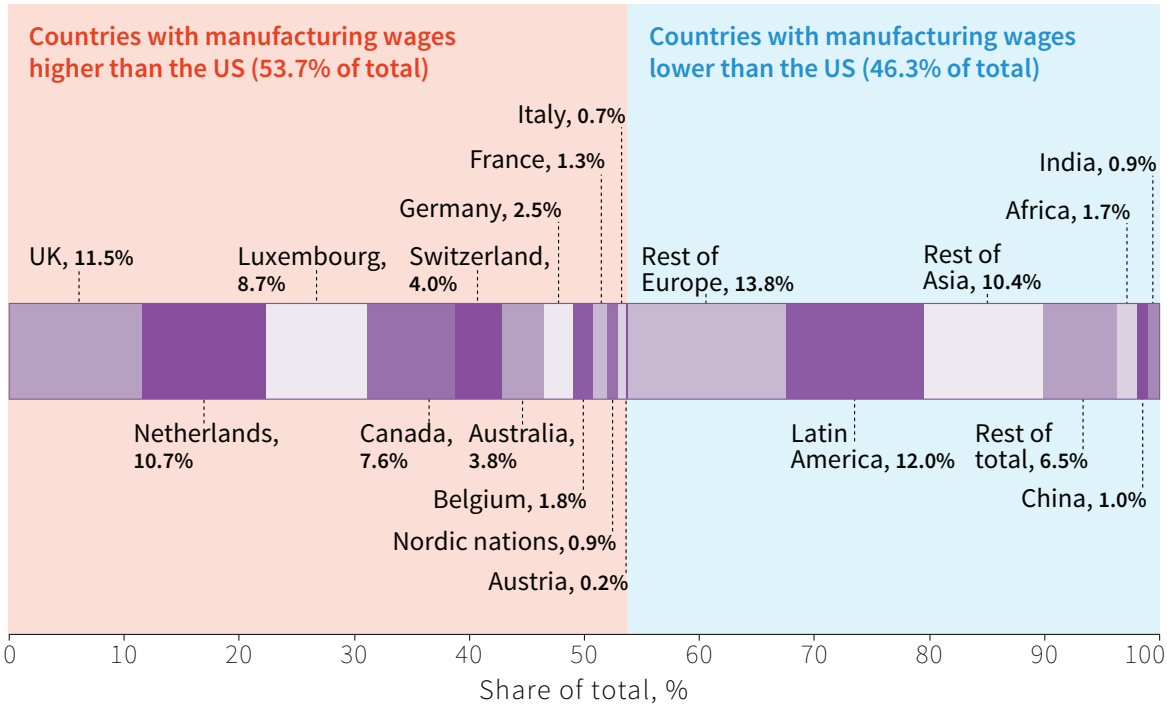


Figure 16.10 *Investment by US firms in other countries according to whether wages are lower or higher than in the US (2001-2012).*

Source: *United Nations Conference on Trade and Development. 2014. 'Bilateral FDI Statistics.' Note: Data for US FDI flows abroad. The countries shown to have manufacturing wages higher than the US are those that are classified by the US BLS International Labor Comparisons as having higher hourly compensation in manufacturing than the US on average over the 2005-09 period.*

16.3 GLOBALISATION AND MIGRATION

In the late 19th century, declining transport costs and rising wages made passage to America affordable for millions. Since then, labour migration is probably the dimension of globalisation along which international economic integration has advanced the least. Indeed, labour into and out of some countries is less mobile internationally today than it was in 1913. Figure 16.11 plots immigration into the US as a percentage of the increase in the US population. During the late 19th and early 20th century immigrants accounted for more than half of the increase in the US population, their numbers more than equalling the number of births minus the number of deaths. Restrictive legislation curbed immigration between the wars. Although the contribution of immigrants to the growth in population has been rising again since the second world war, it has not matched the growth before 1914.

There were relatively few barriers to immigration in the late 19th century. Today migrants without appropriate documentation may be deported or imprisoned. This meant that when Europe was experiencing its population boom, as death rates fell sharply and birth rates fell only with a lag, it was able to ship its surplus population to the relatively empty lands of what the 15th century explorer Amerigo Vespucci had named the "New World" in America. Today's lower-income countries are not so fortunate. Immigration barriers were already in place during the late 19th century, but they became much stricter during and after the first world war, and rich countries retain strict immigration barriers to this day.

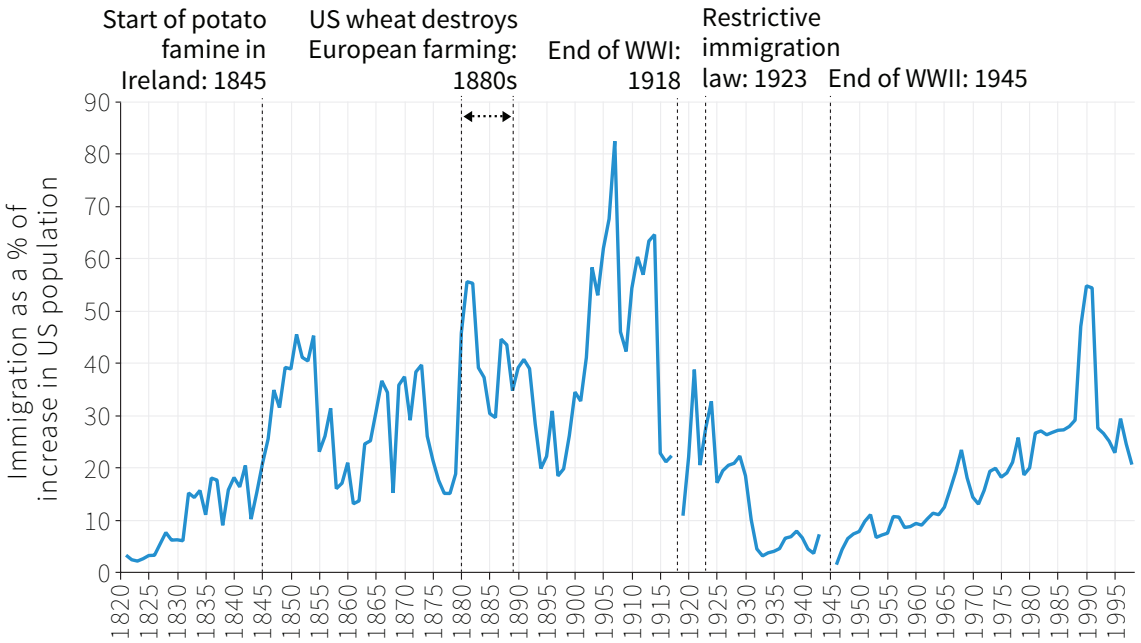


Figure 16.11 *Immigration into the US as a percentage of the change in US population (1820-1998).*

Source: *Carter, Susan B., Michael R. Haines, Richard Sutch, and Scott Sigmund Gartner, eds. 2006. Historical Statistics of the United States: Earliest Times to the Present. New York: Cambridge University Press.*

Thus the movement of goods and finance between countries is easier, and greater in magnitude, than the movement of people. Sending your money or your goods to some distant economy is much easier than sending yourself: you might have to learn an entirely new language or culture. This is one reason why, for labour, there is nothing equivalent to the reduction in price gaps for goods that we discussed above. There is no tendency of wages in different countries around the world to become more similar.

Figure 16.12 shows the trends in wages paid to manufacturing workers expressed as a ratio of the wages of US manufacturing workers. It indicates, for example, that in the late 1970s, workers in Finland were paid 80% of the wage of US workers; but by 2005-09 they were paid more than 30% more.

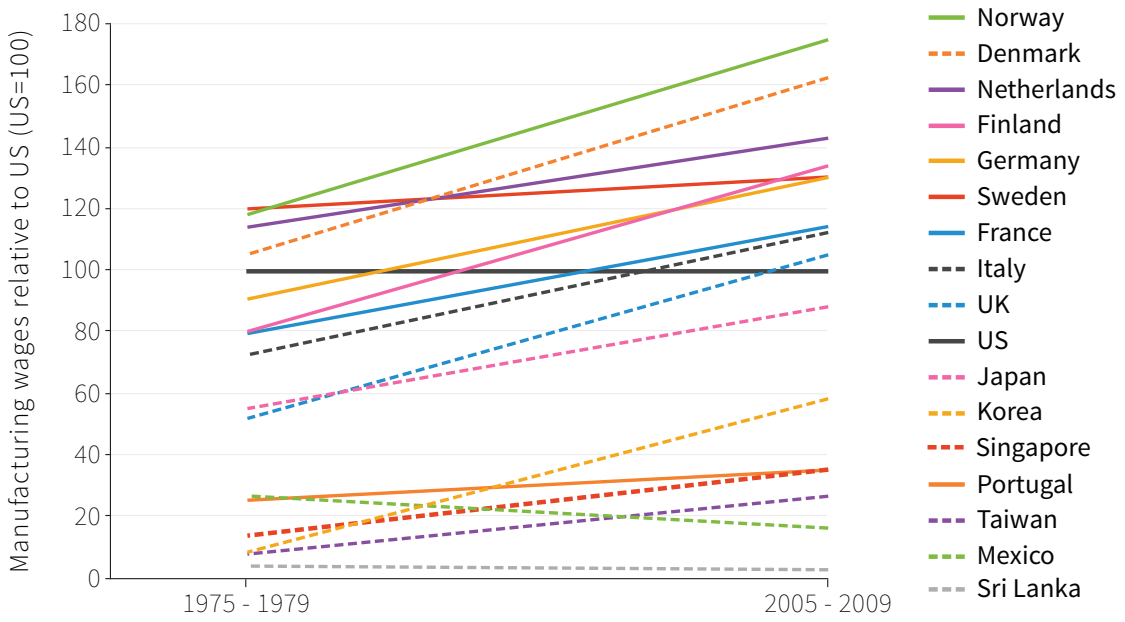


Figure 16.12 *Wages of workers in manufacturing relative to the US (1975-79 and 2005-09).*

Source: *US Bureau of Labor Statistics. 2015. 'International Labor Comparisons.' Note: Data is for hourly compensation costs in manufacturing, which includes total hourly direct pay (pre-tax), employer social insurance expenditures, and labour-related taxes. National currency data converted into US dollars at the average daily exchange rate for the reference year.*

Three things that we can see in Figure 16.12:

- Like Finland, many other European countries caught up with the US in manufacturing wages, in some cases surpassing them by more than 60% (Norway and Denmark).
- Wages in two Asian economies, South Korea and Japan, rapidly converged towards the US wage level.
- A number of low-wage countries (for example, Sri Lanka) remained far behind, some falling even further behind (Mexico).

In conclusion, there was a dramatic increase in the integration of the world economy during the 19th century. This was marked by increasing volumes of trade and corresponding reduction in price gaps, as well as the movement of capital. This was followed by a brief period of deglobalisation during the Great Depression and the second world war, and renewed globalisation afterwards, especially since the 1990s. These three waves of globalisation, deglobalisation and reglobalisation are equivalent to those in Figure 16.4.

Trade costs (and barriers to the mobility of capital and labour) fell in the 19th century, largely as a result of steam-driven transportation technologies. They rose again in the interwar period, largely due to government intervention—taxes and other barriers to trade, capital controls and immigration restrictions—and fell again in the late 20th century, as a result of more liberal policies and technological change.

National boundaries, however, have continued to be important barriers to the global integration of labour markets. National states continue to play an important role in affecting the flows of goods and investment across national boundaries.

16.4 COMPARATIVE ADVANTAGE, SPECIALISATION AND THE GAINS FROM TRADE

Specialisation and trade

The result of this process of integration means that today virtually all nations are part of a global economy characterised by:

- *Specialisation*: particular locations specialise in the production of distinct goods, and
- *Trade*: these goods are then exchanged with other locations specialising in other goods.

Machine tools (such as precision cutting tools for car factories) produced in southern Germany and computers produced in coastal south China are sold throughout the world. Producers of these commodities put food on their table grown in the US or Ukraine and wear shirts made in Mauritius.

As these examples show, trade and specialisation are two sides of the same process: each provides the conditions necessary for the other. In the absence of trade, machine tool workers in Stuttgart could not eat bread made from grain grown in Ukraine or the US and wear clothes made in Mauritius. If they had to provide for themselves, some of them would be farmers or clothing workers. In the absence of specialisation, there would be nothing to trade.

SPECIALISATION

This takes place when a country or some other entity produces a more narrow range of goods and services than it consumes, acquiring the goods and services that it does not produce by trade.

Why do southern Germans specialise in producing machine tools, high-end automobiles and other manufactured goods, while the southern coast of China is the world centre of computer production, Mauritians produce shirts, and the residents of Kansas in the US midwest grow grain? Some possible answers:

- Ukraine and Kansas have the best climate and soil for growing grain.
- Mauritius has few resources other than its population's working capacity, and it takes a lot of labour to produce clothing.

- Germany got a head start in producing manufactured goods that require a high level of skill, and maintains its advantage through economies of scale and distinctive German institutions, both of which ensure high levels of productivity from its workforce.

But what about China's dominance in computer production? China did not benefit from a head start in this industry; it displaced the early leader—the US—only recently.

There is no entirely adequate explanation of global specialisation. Economic research has identified two influences:

- *Comparative advantage*: The fact that some places are better at producing some goods compared to others. Midwestern Americans tend to be grain farmers rather than shirtmakers, for example.
- *Head starts, and economies of scale and external economies from co-location*: In Germany, large firms produce at lower unit costs and, due to its cost advantages, the German machine tool industry is highly competitive, sustaining the large scale of production. Medium-sized firms also benefit from a large local pool of workers with the required skills and experience in the industry. Firms also share information and develop common industry standards for components, and they stimulate research in the region, from which they benefit.

These external economies from similar firms locating in the same area result in lower costs for the entire industry and are sometimes termed *economies of agglomeration* to distinguish them from economies of scale, which refer to a particular firm.

ECONOMIES OF AGGLOMERATION

The cost reductions that firms may enjoy when they are located close to other firms in the same or related industries.

- Don't confuse them with economies of scale or economies of scope, which apply to a single firm as it grows.

These influences on specialisation—comparative advantage, economies of scale and agglomeration—mean we can say that specialisation results from a combination of history, geography, population, engineering and institutions.

We will focus on how geography, population and other factors determine a country's comparative advantage.

Comparative advantage

Imagine two families who live on two islands like the ones in Unit 10, except that at the start they do not trade any goods between them. Greta and her family live on Wheat Island, and Carlos and his family on Apple Island. The land on each island can

be used for growing both wheat and apples, and the families on each island consume both wheat and apples. From now on, when we refer to *Greta* and *Carlos*, we mean their respective families.

Greta is lucky: Wheat Island has better soil for both crops. If Greta devotes a year's labour to growing apples, she produces more apples than Carlos, even if Carlos worked all year on Apple Island. And a year of labour on Wheat Island produces more wheat than a year of labour on Apple Island.

You may wonder how the islands got their names, and why Greta would want to buy apples from Carlos? The answer is that, while Greta can produce more of either crop in a year, she can produce much more wheat than Carlos in year *but only a few more apples*. This is because Apple Island has worse soil than Wheat Island, but also has more hills. It is worse for producing both crops, but it is especially ill suited for producing wheat. We say that Greta's absolute advantage—how much she can produce if that's all she did—is greater in both crops; but her comparative advantage—the thing that she is comparatively good at—is wheat. Carlos is relatively good at producing apples and that's why Greta will buy apples from Carlos.

When people can exchange goods among one another, people and nations specialise in things in which they have a comparative advantage, and then get the other goods they need by trade. As a result, even those who have an absolute advantage in nothing at all will specialise in the thing at which they are *least bad*, and get the other goods they consume by exchange. Similarly, people who are better at producing everything will not produce everything, but instead will specialise in producing the thing at which they are *comparatively best*, while importing other goods. Both Greta and Carlos can benefit from specialisation and trade.

COMPARATIVE AND ABSOLUTE ADVANTAGE

- A country has *absolute advantage* in the production of a good if the inputs it uses to produce this good are less than in some other country.
- A country has *comparative advantage* compared to some other country in the production of the good for which it has the greatest absolute advantage, or least productivity disadvantage.

To see how this works we look at The left-hand panel of Figure 16.13, where the feasible frontier shows all the possible combinations of wheat and apples that Carlos can produce in a year. First, we ask how many apples Carlos can produce if he produces only apples. This is shown on the horizontal axis by point *A*; Carlos can produce 10,000 apples. Similarly if Carlos produces only wheat, he can produce 4,000 bushels of wheat, as shown by point B on the vertical axis. The line that joins points *A* and *B* is the feasible production frontier for Carlos. It shows all the combinations of wheat and apples that can be produced by Carlos in a year. Carlos can choose to produce any combination on (or inside) the frontier. For example, he could produce 2,000 bushels of wheat and 5,000 apples (as shown by point *C*).

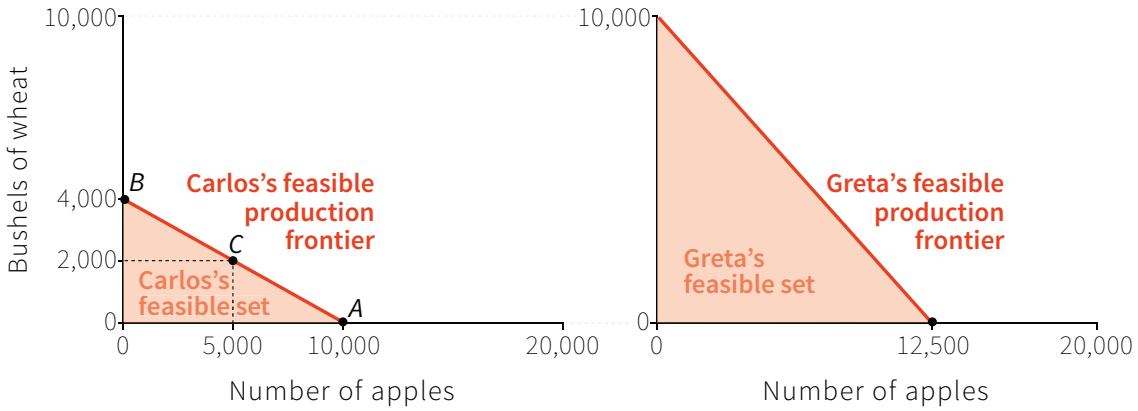


Figure 16.13 *Carlos's (Apple Island's) and Greta's (Wheat Island's) feasible production frontiers.*

Diversification in the absence of trade

The feasible production frontier was first introduced when we met Angela, the self-sufficient farmer in Unit 3. Angela's feasible production frontier curved outward (it was concave), because the amount of grain Angela could produce per hour fell the more time she spent farming. In this example, we assume that the feasible production frontier is a straight line. In other words, it always takes Carlos the same amount of time to produce an apple, regardless of how many apples have already been produced. Angela had diminishing marginal productivity of labour, but Carlos has *constant marginal productivity of labour*. It is feasible for Carlos to produce anywhere between the origin and the feasible production frontier, as shown by the shaded area in the left-hand panel of Figure 16.13. The feasible set shows all the combinations of wheat and apples that Carlos can produce in a year, given the state of technology. Carlos will always choose to produce on the frontier: for all points inside the frontier he could produce more of at least one good, without reducing production of the other good.

We show the feasible production frontier for Greta in the right-hand panel of Figure 16.13. Greta can produce more of both goods in a year than Carlos can. She can produce 12,500 apples or 10,000 bushels of wheat. When one island can produce more of a good using the same labour input (imagine Carlos and Greta's families are the same size), they are said to have an absolute advantage in producing that good. In our example, Wheat Island has an absolute advantage in producing both goods. We can see that this means that Greta's feasible set is much larger than Carlos's; she can produce a number of combinations of wheat and apples that are unavailable to Carlos.

GREAT ECONOMISTS

DAVID RICARDO

David Ricardo (1772-1823) developed the theory of comparative advantage. He was also the first economist to warn that a rapidly growing capitalist economy would confront the limits of its natural environment.

The son of a successful stockbroker and the third of 17 children, Ricardo grew up in London and eloped at the age of 21, which led to a long period of estrangement from his parents. He went on to build a huge fortune through trading in stocks before interesting himself in political economy. He entered parliament (by purchasing a seat, which then was possible) where, besides his contributions on economic questions, he favoured liberal social causes such as religious tolerance, freedom of speech and opposition to slavery.

Ricardo's central contribution was an analysis of the principles of production and distribution in a growing capitalist economy with a large agrarian sector. The Ricardian model, which he developed and which came to dominate British economic thought for much of the next 50 years, was laid out in *An Essay on Profits*, published in 1815. Agricultural production relied on three inputs, labour, capital and land, and as production and population expanded, either existing land had to be farmed more intensively with greater doses of capital and labour, or less fertile plots had to be brought into production.

Drawing on ideas of diminishing returns, he explained how this would lead to a squeeze on profits and the eventual stagnation of the economy. Like Thomas Malthus, whose ideas we studied in Unit 2, he reasoned that wages could not be below subsistence. As farming expanded to less good land, the price of food and hence wages would have to increase. A result would be that profits (which Ricardo presumed would be invested) would fall. Rents (presumed to be spent on luxuries) would increase due to the growing scarcity of land. The result would be the eventual slowdown and stagnation of the economy.

Ricardo therefore advocated a repeal of tariffs on the import of grains (known as the *Corn Laws)*, which his friend Malthus defended. Ricardo reasoned that if Britain could acquire more of its food from the US and elsewhere, then paying workers a subsistence wage would cost less to the employers, raising the rate of profit and investment. Importing grain rather than growing it in Britain would make land less scarce and therefore limit the landlord's share of total output. The result, according to Ricardo, would be continued growth rather than stagnation.

His greatest work, *Principles of Political Economy and Taxation* (published in 1817), introduced the labour theory of value, later used by Karl Marx. This theory holds that the value of goods is proportional to the amount of labour required, directly or indirectly, in their production. Wassily Leontief (1906-1999), an economist, devised a way that these values could be calculated (see W*hen Economists Disagree: Heckscher, Ohlin, and the Leontief Paradox* in the next section).

In *Principles*, Ricardo discovered the principle of comparative advantage, recognising that two countries could trade to the mutual advantage of each, even if one of them was absolutely better at producing all goods.

Ricardo is not as famous an economist as Smith, Malthus, Mill or Marx; but he is greatly respected for the theory of comparative advantage, and his method of thinking thoroughly through the dynamics of an abstract system as a guide to economic understanding. In this way, he is a very modern Great Economist.

How do we know how much Greta and Carlos consume of each good? Recall Alexei's decision of how many hours to study in Unit 3, and Angela's decision about how hard to work her landlord's land in Unit 5; maximisation of utility motivated both decisions. In each case, we have modelled utility using indifference curves. The further the indifference curve is from the origin, the higher the level of utility it represents. Carlos and Greta will aim to get onto the highest indifference curve possible given the constraint of their feasible production frontier. In our simple example, the feasible production frontier is also the feasible consumption frontier, because each family spends time producing only wheat and apples, and can consume only the amount they produce.

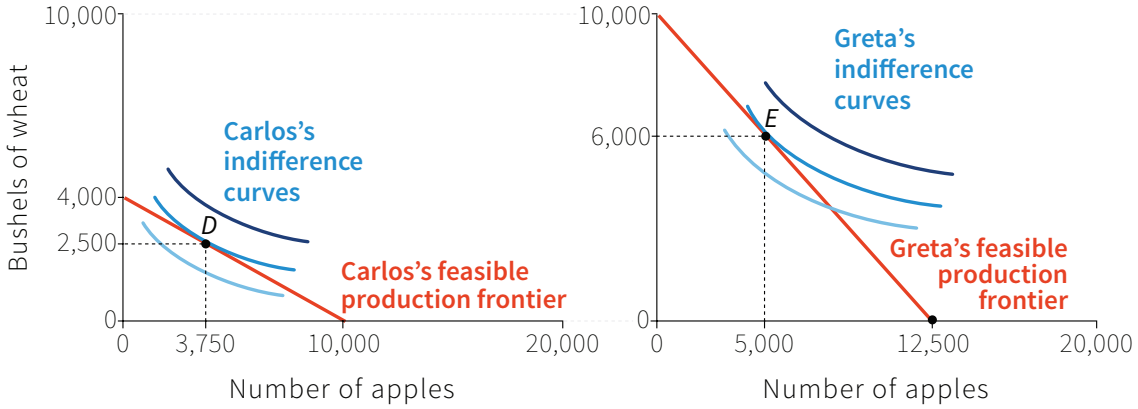


Figure 16.14 *Carlos's (Apple Island's) and Greta's (Wheat Island's)utility-maximising choices of consumption.*

The left-hand panel of Figure 16.14 shows Carlos's feasible consumption frontier. We now add Carlos's indifference curves. The shape of the indifference curves represents Carlos's preferences over wheat and apples. The highest indifference curve he can attain will be the one that is tangent to his feasible consumption frontier. He will therefore choose to consume 2,500 bushels of wheat a year and 3,750 apples, as shown by point *D* in the right-hand panel of Figure 16.14. Greta's superior productivity means she can consume more of both goods than Carlos. Given her preferences, which are the same as Carlos's (the indifference curves are the same shape), she consumes 6,000 bushels of wheat a year and 5,000 apples, as shown by point *E* in the right-hand panel of Figure 16.14.

Trade and specialisation on Wheat and Apple Islands

At this point, we have observed Wheat and Apple Islands prior to trade. What will happen when Greta and Carlos and their families are able to trade? The decision to trade could be made for a number of reasons, such as the development of a new technology (maybe a boat) or the removal of barriers to trade (perhaps the end of a feud between the two islands).

Does the absolute advantage of Wheat Island in the production of each good mean that there are no potential gains from trade between the islands? No, because it is the relative, not the absolute, cost of producing the two goods that matters for mutually beneficial trade.

We will show that both Carlos and Greta gain when one island specialises in the production of wheat and the other specialises in the production of apples. The decision of the good in which to specialise depends on the relative prices of the two goods when Carlos and Greta are unable to trade with one another. The prices depend on the cost of producing the good, and this depends on how productive are the workers. The relative prices depend on how much of each good a single worker can produce in a year.

We assume that there are 10 members in both Carlos and Greta's families, and that they are all equally productive. Carlos's family can produce 4,000 bushels of wheat a year, or 10,000 apples. In order to produce one more bushel of wheat Carlos's family has to produce 2.5 fewer apples, so the marginal rate of transformation between bushels of wheat and apples is 2.5. Since it takes the same amount of inputs (land and labour) to produce one bushel of wheat as it does to produce 2.5 apples, a bushel of wheat will cost the same as 2.5 apples. Thus the relative price of wheat to apples will be 2.5. Following the same logic, the relative price of apples to wheat is *400/1,000 = 0.4*.

Greta's family is more productive in producing both goods. A member of Greta's family can produce 1,000 bushels of wheat in a year, or 1,250 apples. The relative prices of wheat and apples on Wheat Island are therefore 1.25 and 0.8.

Wheat Island has a comparative advantage in producing wheat: the island can produce wheat relatively more cheaply than Apple Island can; for each extra unit of wheat Greta produces she only has to give up 1.25 apples (compared to 2.5 on Apple Island). The relative price of apples is simply the inverse of the relative price of wheat so, if Wheat Island has a comparative advantage in producing wheat, then Apple Island will have a comparative advantage in producing apples. Figure 16.15 summarises the key numbers from the example. The relative prices of the good for which each island has a comparative advantage are shaded in dark blue.

	Apple Island (Carlos)	Wheat Island (Greta)
Bushels of wheat produced by each family member in one year	400	1,000
Number of apples produced by each family member in one year	1,000	1,250
Relative price of wheat	1,000/400 = 2.5	1,250/1,000 = 1.25
Relative price of apples	400/1,000 = 0.4	1,000/1,250 = 0.8

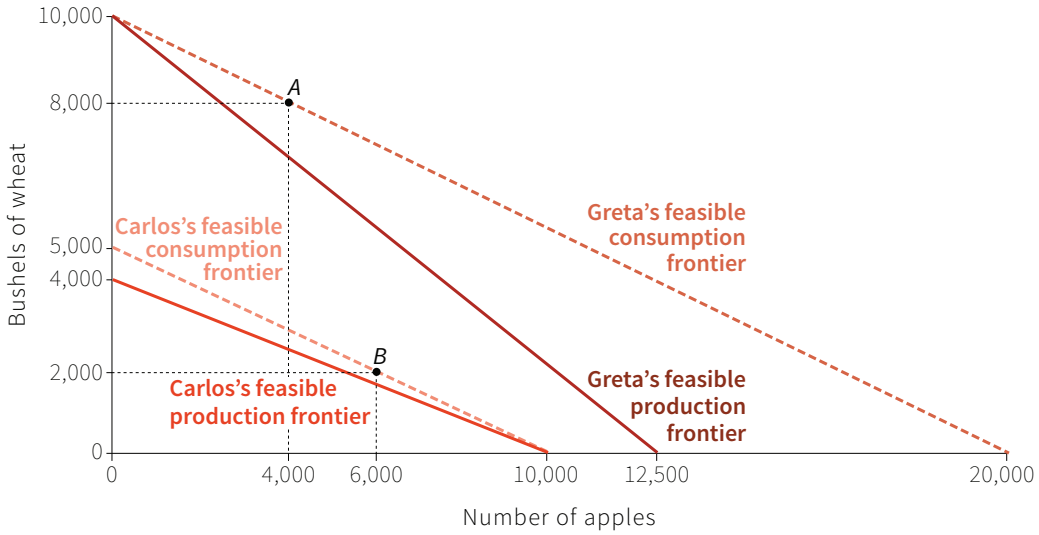
Figure 16.15 *An island has a comparative advantage in producing a good when it is relatively cheaper in their economy (in the absence of trade).*

Gains from trade

When there is no trade (we say the economies are closed), the feasible production frontier is also the feasible consumption frontier. Figure 16.14 showed that when the economies are closed Carlos produces 2,500 bushels of wheat and 3,750 apples, while Greta produces 6,000 bushels of wheat and 5,000 apples. Total production between the two countries is therefore *2,500 + 6,000 = 8,500* bushels of wheat and *3,750 + 5,000 = 8,750* apples. When countries specialise, however, Greta will produce 10,000 bushels of wheat and Carlos will produce 10,000 apples, so there is more of both goods. As long as they can trade, so that both families get to consume both goods, they can both be made better off.

If we assume that there are no trade costs, then the relative price of wheat and apples will have to be the same in the two countries. What will the new price be? From Carlos's point of view the supply of wheat has increased more than the supply of apples, so the price of wheat relative to apples will go down for him and his family to something less than 2.5. Equally, from Greta's point of view the supply of wheat has increased less than the supply of apples, so the relative price of wheat will go up for her to something higher than 1.25. This shows that the relative price when the economies are open, and trading, will be between the prices experienced by the two economies when they are closed.

Let us suppose that the relative price of wheat is 2 (a price between 1.25 and 2.5). We can now redraw the feasible consumption frontiers for the two economies in Figure 16.16, because the slope of the frontier is equal to the relative price. For Greta the slope gets flatter because wheat has become relatively more expensive (rising from 1.25 to 2), while for Carlos the slope gets steeper because wheat has become relatively cheaper (falling from 2.5 to 2). But because both countries are now specialising in the good in which they have a comparative advantage, the new consumption frontiers are above their production frontiers. We can see that specialisation and international trade have led to an increase in the size of the feasible consumption set for both countries, even though Greta has an absolute advantage in producing both goods. If we look back to Figure 16.14 we can see that any expansion of their feasible sets make it possible for both Carlos and Greta to reach a higher level of utility; trade has been mutually beneficial.



Consumption after specialisation and trade

The dotted red lines show the outward shift of the feasible consumption frontiers due to specialisation and trade. We assume that the relative price of wheat after specialisation and trade is 2 (a price in between 1.25 and 2.5).

Figure 16.16 *The effect of trade and specialisation on Carlos and Greta's feasible consumption frontiers.*

Specialisation has enlarged the feasible consumption set of both in the same way that borrowing and investing increased the feasible consumption set of Marco in Unit 11. By investing, Marco *specialised* in having income in the future, and then by borrowing he *imported* some of his future income to the present so that he could consume in both periods.

How much will Greta and Carlos trade? One thing we know is that the amount of wheat Greta exports has to be exactly equal to the amount of wheat Carlos imports (it is the same wheat!), and the same is true for the apples that Carlos exports, and Greta imports. In Figure 16.16, we show two points indicating consumption for each resulting after specialisation and trade. Points *A* and *B* in Figure 16.16 are a jointly feasible outcome. Carlos specialises in apples and exports (*10,000 – 6,000 = 4,000*) apples to Greta who specialises in wheat and exports (*10,000 – 8,000 = 2,000*) bushels of wheat to Carlos.

The relative price determines the extent to which trade increases the feasible set of each island. This, in turn, depends on how the price is determined. Suppose that Greta can unilaterally determine the price. To increase her gains from trade, Greta will choose a price that increases the amount of apples she receives for each bushel of wheat she sells to Carlos. If we assume that she has chosen a price of wheat of 2.25, then how does this affect the expansion of the feasible sets?

Figure 16.17 starts with the same feasible production frontiers as Figure 16.16. After trade, Greta dictates the relative price of wheat to be 2.25. Trade still shifts out the feasible sets of both Carlos and Greta. We can see that the gains from trade are not as equally distributed as in Figure 16.16. Greta's feasible set has shifted out a lot, and Carlos's has not shifted very far. This means trade and specialisation will increase the utility of both Carlos and Greta, but will increase Greta's utility by more.

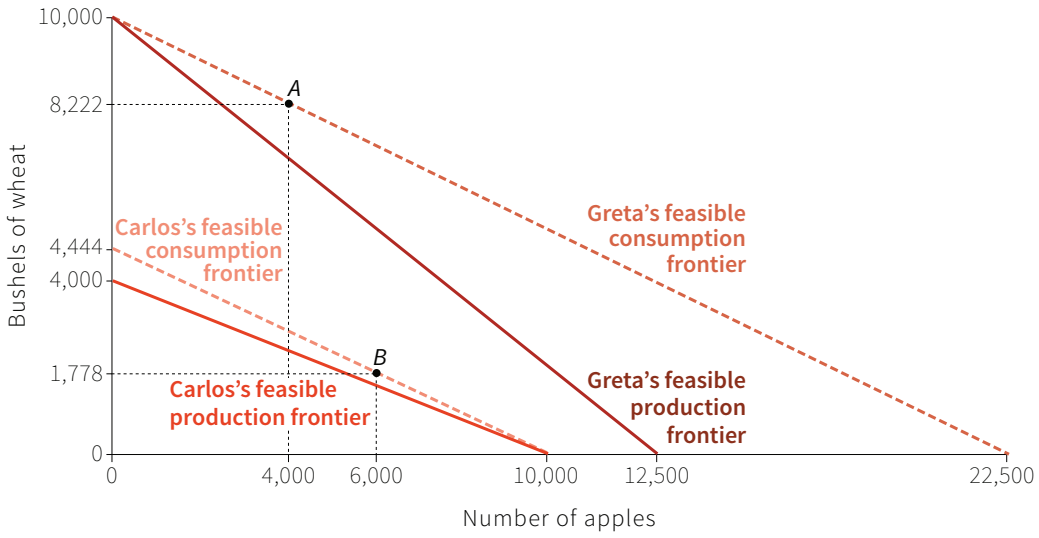


Figure 16.17 *The effect of trade and specialisation on the feasible consumption frontiers for Carlos and Greta, when Greta is able to dictate the price.*

In Figure 16.17, as in Figure 16.16, we show a possible trade between the two of them, this time using the new price dictated by Greta. To get the 4,000 apples that she wants, she now has to give up only 1,778 bushels of wheat (at the old price, she had to give up 2,000). Carlos, meanwhile, (at point *B*) is worse off than before. If he still wants to consume 6,000 apples, he can get only 1,778 bushels of wheat in return for exporting 4,000 apples rather than the 2,000 he got before.

In our example, specialisation and trade has been mutually beneficial (even when the gains were not equally distributed); it has raised living standards for both Greta and Carlos. In other words, there were no losers from specialisation and trade.

Of course, if Greta could set any price she wished, she could have set an even higher price. If she set the price at 2.5 apples per bushel, she would eliminate Carlos' gains from trade entirely. At this price Carlos would be equally well off if he produced his own wheat or bought wheat from Greta.

When the people of a country benefit from influencing the price in their favour we say they have *bargaining power*, which we now look at in more detail.

DISCUSS 16.4: COMPARATIVE ADVANTAGE

Suppose that each worker in Germany can produce three cars or two televisions, and there are just four workers in Germany. The only other country in the world is Turkey, where each worker can produce two cars or three televisions. Assume that Turkey also has four workers.

1. Draw the feasible production frontier for each country. What is the relative price of cars in each country without international trade?
2. Suppose that, in the absence of trade, Germany consumes nine cars and two televisions and Turkey consumes two cars and nine televisions. Mark these consumption points without trade as G and T, respectively. Draw the feasible consumption frontier in each country without trade. Comment on the relationship between the production and consumption frontiers you have drawn for each country.
3. Now suppose Germany and Turkey start trading. What is the range of possible values for the world relative price of cars? If the world relative price of cars is PC / PTV = 1, in which good will each country specialise?
4. Now use the world relative price given above to draw the feasible consumption frontier in each country in the figures you have drawn above. Use these figures to discuss whether each country gains from trade.

DISCUSS 16.5: POWER AND BARGAINING

Use what you learned in Unit 4 about how people play the ultimatum game: How do you think Carlos would react to a price of 2.5 apples per bushel of wheat?

16.5 WINNERS AND LOSERS

Carlos and Greta both benefit from trade, so why are imports and exports often controversial? Unlike our story, in the real world there are almost always winners and losers. The processes of specialisation and exchange affect regions, industries and household types differently. Had the bakers and shoppers of Genoa known that cheap grain was aboard the *Manila* they would have cheered her into port, while local farmers secretly prayed for a shipwreck.

Nations are composed of people with differing economic interests. They are not like our islands with only Greta, Carlos and their (we imagine) happy families.

To think about winners and losers from trade, we use a model based on comparative advantage between two stylised countries, which we call Germany and China. Germany is an advanced economy with a long tradition of manufacturing high-quality goods. China is less developed, but has become the world's second-largest economy by focusing on exporting manufactured goods. Let us imagine that Germany and China produce only two goods: machine tools (such as precision cutting tools for a car factory) and consumer electronics (like MP3 players, personal computers and TVs). Furthermore, we assume (more realistically this time) that Germany has an absolute advantage in producing both goods, and a comparative advantage in producing machine tools.

Germany's comparative advantage arises because machine tool production is capital-intensive, requires lots of expensive machinery, and capital is relatively abundant in Germany. In contrast, China has a comparative advantage in consumer electronics production, which is more labour-intensive, and China has an abundance of labour relative to capital. For example, LG Electronics factories in China are staffed by tens of thousands of assembly-line workers. Given these assumptions, when the economies are able to trade with one another, Germany will specialise in the production of machine tools, and China will specialise in the production of consumer electronics.

We know the effect of opening trade between Germany and China in machine tools and consumer electronics:

- It increases the consumption possibility set for both countries.
- Conflicts of interest emerge between the countries, and within each country.

The relative price of the two goods affects how the gains from trade are divided between the countries. On Wheat and Apple Islands the usual forces of demand and supply affect the relative price, but the balance of bargaining power between the two families affects the price too. For Germany and China, and all other countries in the real world, relative prices are subject to the same forces.

In Unit 14, for example, we investigated the macroeconomic consequences of oil price shocks. But what caused the increase in the relative price of oil?

- *The first and second oil shocks (1970s)*: Relative price increases were due to political developments in the Middle East and the ability of oil producers to exert monopoly power through a cartel. The exercise of monopoly power by producers shifted the supply curve upward.
- *The third oil shock (2000s)*: The growth of China and other emerging economies caused a large increase in global demand. The global demand curve for oil shifted to the right.

The beneficiaries of a rise in relative price are the inhabitants of the country that specialises in producing that product. But do all citizens benefit? There may be a conflict of interest within the country, which we did not see in Apple or Wheat Island. In that model, each family shared the products available. But not everyone in a country is the same. For example, some people have only their labour to sell. Others have accumulated wealth, which they can use to invest in firms.

In the example of Germany and China, after trade Germany specialises in producing machine tools and China specialises in consumer electronics. Trade and specialisation mean that resources shift from one industry to another. Workers previously employed in electronics in Germany must try to find work in machine tools firms. Similarly, in China, employment will expand in consumer electronics production.

The increase in production of machine tools in Germany increases the demand for the factor of production used intensively in that industry: capital. In China, it increases demand for labour instead.

- *The winners in Germany*: The owners of capital benefit more from trade than workers, because capital becomes relatively scarce as production of machine tools rises. We would predict a rise in inequality.
- *The winners in China*: Workers are in higher demand as consumer electronics production expands. Wages rise as firms compete for workers. As we have seen in Unit 6, lower unemployment lowers the cost of job loss, and firms raise wages. This will reduce inequality. Workers benefit more from trade than the owners of capital.

Trade and specialisation in Germany involves transferring labour and capital from electronics production to machine tool production. Think of what happens when one unit of capital (think of a factory) shifts from electronics to machine tool production. An electronics factory closes, laying off *X* workers, and a machine tools factory opens, hiring *Y* workers. Which is bigger, *X* or *Y*?

The answer: *X* is bigger than *Y*, since one unit of capital provides the tools and equipment necessary to employ more workers in electronics than in machine tool production (since electronics is relatively labour-intensive). Thus, when capital shifts from electronics production to machine tools production, there is a net loss of jobs.

In this case, German workers are losing out, and German employers are winning: workers are working for lower wages, and profits rise. The effect of imports of labour-intensive electronics and the shift in German production to less labour-demanding goods (machine tools) is that employers capture most of the gains from trade. As consumers of electronics, both employers and workers benefit. This is an example of a general principle about who benefits from international trade: The owners of relatively scarce factors of production (German labour in our example) lose from specialisation and trade, and the owners of relatively abundant factors (owners of capital in Germany) gain.

The reasoning behind this principle is as follows:

- Factors that are relatively scarce in their own countries, compared with in the rest of the world, are relatively expensive when there is no trade. When their economies start trading with the rest of the world it drags their price down towards the world average, because they are effectively competing with their abundant counterparts in the rest of the world.
- Relatively abundant factors are relatively cheap, and gain when trade raises their price towards the world average.

So, in Germany, workers are relatively scarce and lose from trade, while employers gain; in China workers are relatively abundant and gain from trade, while employers lose. The key to understanding this is to focus on the change in relative scarcity once labour and capital can flow across economies embodied in traded goods and services.

Figure 16.18 illustrates the two dimensions of conflict arising from international trade.

On the left we have German and Chinese economies, with limited specialisation and trade. The economies are normalised to a size of one, and the numbers in the pies show both the proportion and size (in brackets) of the slice of the pie that accrue to workers (red) and the owners of capital (blue). The pies on the right show the German and Chinese economies with greater specialisation and trade. The gains from specialisation and trade are clear from the fact that the pies on the right are larger.

The size of the German economy has increased by 30% and the size of the Chinese economy has increased by 40%. The prices at which they have traded have resulted in China securing more of the gains from trade.

But notice too that China's shift into labour-intensive electronics has raised labour's share of China's larger pie, and reduced the share of profits. Both capital and labour in China are, however, better off with higher specialisation and trade, as the absolute size of the slices going to workers and the owners of capital have both increased (*0.5 < 0.84 and 0.5 < 0.56*).

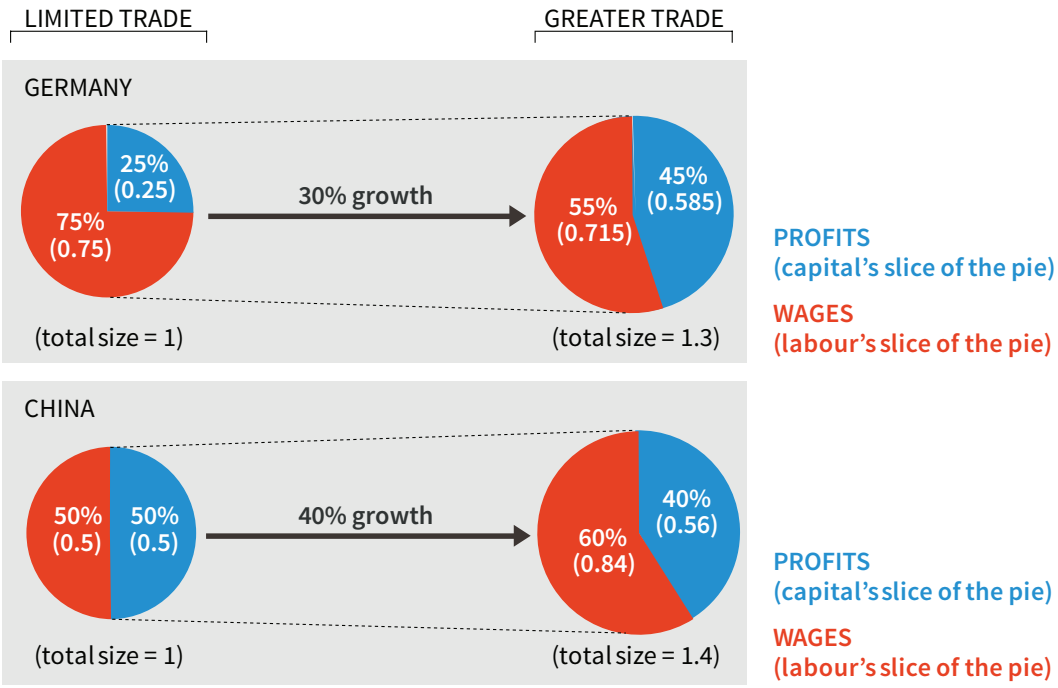


Figure 16.18 *The winners and losers from trade between Germany and China.*

The story is different in Germany. The owners of capital goods (employers) now have a larger slice of Germany's larger pie, but the German workers' slice is not only proportionally smaller (*75% > 55%*) but also smaller in absolute size (*0.75 > 0.715*). German workers are the losers. German employers, Chinese employers, and Chinese workers are all winners.

The same logic would continue to apply if we considered other factors of production. For example, consider two industries that require employees with different levels of skills and education: a skill-intensive industry (information technology) and a non skill-intensive industry (consumer electronics assembly). If a rich economy, relatively abundant in skilled labour, starts trading with a poor, unskilled, labour-abundant country, then unskilled workers in rich countries (and skilled workers in poor countries) will lose, while skilled workers in rich countries (and unskilled workers in poor countries) will gain. You might think that this would affect the way

that different groups viewed trade. Indeed there is considerable survey evidence that unskilled workers in rich countries are more protectionist than skilled workers, but unskilled workers in poor countries are more in favour of trade than skilled workers.

The example of Germany and China shown in this section does not only have relevance for the wave of globalisation after 1945. One hundred years ago, when Eli Heckscher and Bertil Ohlin, two Swedish economists, were working on better understanding patterns of specialisation and trade, they were motivated by the globalisation of the late 19th century. The difference between then and now is simply the factors of production involved. While our example of Germany and China focused on capital- and labour-intensive manufactured goods, the globalisation of the late 19th century involved the exchange of land-intensive agricultural goods (food and raw materials such as cotton) for labour-intensive manufactures.

The agricultural goods were exported by land-abundant (and labour-scarce) countries such as the US, Canada, Australia, Argentina and Russia; the manufactures were exported by labour-abundant (and land-scarce) countries in northwest Europe, such as Britain, France and Germany. In this context the big losers were European landowners and workers in land-abundant regions; the big winners were European workers and the owners of land in the New World and other land-abundant economies. In Unit 2 we saw that workers in England gained economically, relative to landowners, from the middle of the 19th century onwards.

The same happened in other land-scarce, labour-abundant societies in Europe and elsewhere (for example, Japan). Meanwhile, the ratio of land rents to wages rose strongly in land-abundant, labour-scarce regions: not just the New World economies mentioned earlier, but also in areas such as the Punjab, which was a major exporter of agricultural products.

Not surprisingly, European landowners objected to this, and in countries such as France and Germany they succeeded in getting governments to impose tariffs on agricultural imports. There was thus a political backlash against globalisation. Governments raised trade costs in the form of tariffs to counteract the impact of the fall in other trade costs, notably transportation.

During the 1980s, economists Avinash Dixit, Elhanan Helpman and Paul Krugman, and others, developed models of trade in which trade was not due to differences between countries, but to increasing returns. If, through specialisation, trade allows countries to reap greater economies of scale, this should make trade an even better idea. This “New trade theory”, however, leads to new arguments for protection too. For example, increasing returns means monopoly profits—so perhaps it would be a good thing if your country gets these profits, rather than someone else. Read Paul Krugman’s Nobel lecture, and an earlier paper that he wrote on free trade, to find out more.

WHEN ECONOMISTS DISAGREE

HECKSCHER, OHLIN, AND THE LEONTIEF PARADOX

If countries were identical none would have a comparative advantage in the production of any good, and there would be no reason for them specialise and to exchange goods based on the logic of comparative advantage. Eli Heckscher (1879-1952) and Bertil Ohlin (1899-1979) reasoned that, when accounting for comparative advantage and trade, the key differences between countries were the relative scarcity of land, labour or capital. Canada and the US had abundant land relative to the amount of labour, and hence would specialise in and export agricultural goods. With more capital and less labour than China, Germany would export capital-intensive goods to China.

Wassily Leontief challenged the widely accepted Heckscher-Ohlin theory in 1954. Using a method of input-output analysis that he had invented, he measured the amount of labour and capital goods used in the production of the goods exported from, and imported to, the US. He determined, for example, the amount of labour required:

- To produce a car
- To produce the steel, that went into the car
- To produce the coal, that fired the steel plant, that produced the steel, that went into the car

… and so on.

Based on the Heckscher-Ohlin theory he expected that, because the US was the most capital-abundant country in the world when measured by the stock of machinery, buildings and other capital goods per worker, its exports would be capital-intensive and its imports labour-intensive.

He found the opposite.

For more than 50 years, economists have struggled to resolve this so-called *Leontief paradox*. Leontief speculated that the US might be labour-abundant if instead of simply measuring the quantity of employees we include cultural and organisational factors that support a high level of effective work per employee. While his hypothesis has not yet been adequately tested empirically, it reminds us that culture and institutions may be an essential part of explaining how an economy works. Some recent research is consistent with Leontief's conjecture and shows that his paradox holds in many countries.

DISCUSS 16.6: THE COLLAPSE OF THE SOVIET UNION

In the late 1980s and early 1990s the Soviet Union collapsed. The Soviet Union comprised Russia and some of the countries that now make up central and eastern Europe and central Asia. It was a planned economy, run from Moscow by the Communist party. Following this collapse, countries in the former Soviet Union and elsewhere in the former Soviet bloc—with a total of close to 300 million workers—opened their borders to international trade.

Use the analysis in this section to identify likely winners and losers from this shock to global trade in:

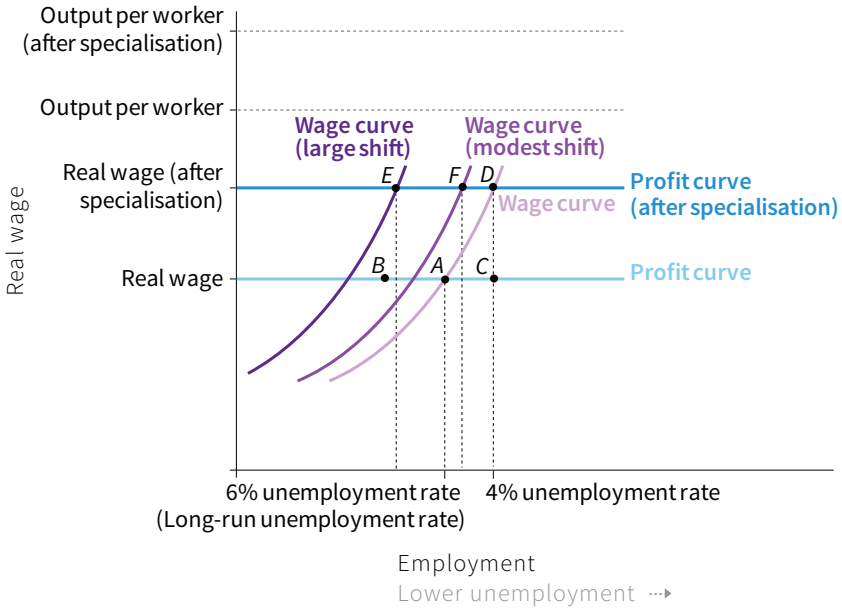
1. Germany
2. The countries of the former Soviet bloc
3. What information do you need to know about these countries?

16.6 WINNERS AND LOSERS IN THE VERY LONG RUN

In our example of Germany and China, the short-run effect of trade was to raise the profits of German employers while depressing the wages of German workers. This would provide German employers with incentives to invest more in building additional capacity to produce machine tools. Our analysis of wages and employment in the long run (in Unit 15) provides a lens for us to study what will happen next. In Figure 16.19 we start with the German wage curve and the profit curve before specialisation and trade with China. The economy starts at point A with unemployment at the long-run rate of 6%.

Specialising in the production of the good in which it has a comparative advantage increases the productivity of German labour (workers have moved from producing electronics to producing machine tools where they are more productive). This shifts up the profit curve and output per worker. So, in this respect, specialisation according to comparative advantage is similar to technological progress.

We can now use Figure 16.19 to trace the effects of this change.



Specialisation and unemployment

The economy starts at point *A* (6%). Following German specialisation in machine tools, the economy moves from point *A* to point *B* and unemployment rises. German firms build new production capacity, expanding the demand for labour and re-employing former electronics workers. The economy moves from point *B* to point *C*, and unemployment falls below its original level, increasing workers' bargaining power. The economy moves to the right and up the wage curve. This process stops at point *D*. The wage curve may also shift if workers demand more unemployment insurance. For example, at point *E*, unemployment is higher than the original long-run rate of 6%. However, if there were only a modest shift in the wage curve, employment would have risen as a result of specialisation, as shown by point *F*.

Figure 16.19 *The effect of specialisation on German long-run unemployment, according to comparative advantage.*

- *Workers producing consumer electronics are laid off*: German consumers are now buying their DVD players from China. Some are hired in producing machine tools, but not all.
- *German machine tool firms are making large profits*: They anticipate this will continue in the future. They build new production capacity.
- *Increased demand for labour increases workers' bargaining power*: The economy moves to the right and up the wage curve.

When will this process stop? When the economy has come to the new intersection of the profit and wage curve at point D. Will the German economy now employ more or fewer workers than before?

The answer depends on the change to the wage curve. In many countries the integration into the world economy was accompanied by unemployment in some sectors of the economy, as well as economic fluctuations due to international price changes. The result was an increase in demand for unemployment insurance, and a strengthening of employment protection and other policies to protect households from shocks to income and employment. Voters supported these policies for the same reason that households seek to smooth consumption. These effects would shift the wage curve up.

In Figure 16.19 you can see that if the wage curve had shifted a lot, the result might have been a reduction in total employment. For example, at point *E* in Figure 16.19 unemployment is higher than the original long-run rate of 6%. But had there been a small shift, employment would have risen, as shown by point F in Figure 16.19.

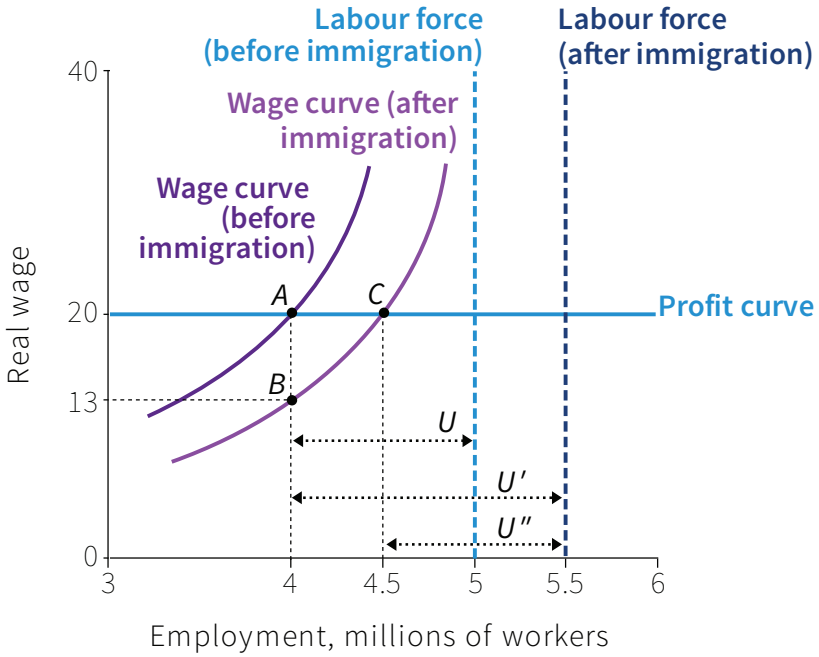
In the era after the second world war, as will see in the next unit, many countries integrated their economies into the world economy and at the same time developed income-smoothing policies commonly termed the welfare state. As we saw in Unit 15, in Nordic countries, for example, trade unions agreed to unimpeded imports. In return they won policies to retrain workers displaced by imports, and which provided support to unemployed workers.

16.7 MIGRATION: GLOBALISATION OF LABOUR

Just as the Italian farmers had not been happy to see the cheap Indian grain being offloaded from the steamer *Manila* in Genoa, workers in North America did not always welcome Europeans in search of a more affluent life, like the 69 passengers sailing west on the *Manila* after leaving Genoa on their way to New York. Immigration hurt unskilled workers in the New World, and where unskilled wages lagged furthest behind average incomes, immigration barriers were raised the most.

So we also saw another type of globalisation backlash during the first period of globalisation in the 19th and early 20th centuries: gradually rising immigration barriers.

To see why opposition to immigration was common among workers in land-abundant economies like the US or Canada, return again to how wages are set by firms (Unit 15). When new people arrive in a nation they are unemployed, so we might expect the first impact of immigration to be that it increases unemployment. This means that immigration also increases the cost of job loss, because the worker who loses a job is now in a larger pool of unemployed workers. Workers have more to fear from losing their jobs, and firms will be able to make employees work effectively at a lower wage.



The initial situation

The economy starts at point *A*, employing 4 million workers at a wage of $20 per hour and a labour force of 5 million.

One million workers are unemployed

This is shown by the distance *U*.

Immigrant workers join the labour force

This increases the labour force from 5 million to 5.5 million workers.

The wage curve shifts downwards

At any level of employment there are now more unemployed workers. The rise in unemployment to 1.5 million is shown by distance *U′*. The threat of job loss is greater and firms can secure effort from the workforce at a lower wage.

Firms lower the wage

It is at point *B* in the figure, with the wage at $13 an hour and employment still at 4 million.

Profits rise

This causes firms to invest and hire more workers, which requires rising wages along the wage curve. The economy moves from point *B* to point *C*.

Employment and wages rise

They rise until they reach the profit curve, meaning profits are normal again. At point *C*, employment is 4.5 million workers, the wage is $20, and unemployment has fallen back to 1 million workers, as shown by distance *U″*.

Figure 16.20 *Immigration and unemployment.*

This is not the end of the story. The firm is now getting work at lower wages, and so is more profitable. As a result it will seek to expand its production. To do this it will invest in new machinery. This will increase labour demand in the rest of the economy, and when the new capacity is ready, the firm will hire more workers.

In Figure 16.20 we show these longer-run effects using the wage curve and the profit curve introduced in Unit 15. We now show the labour force explicitly in the figure. We can use the wage curve to work through the long-run impact of immigration, step by step.

In this story the short-run impact of immigration is bad for existing workers in that country: wages fall and the expected duration of unemployment increases. In the longer run the increased profitability of firms leads to expanded employment that eventually will (if no further changes in the situation take place, like another wave of immigration) restore the real wage and return the economy to its initial rate of unemployment. As a result, incumbent workers are no worse off. Immigrants are likely to be economically better off too—especially if they left their home country because it was difficult to make a living.

However, studies of immigration suggest that the predicted negative effect on workers typically do not occur even in the short run. This is because many immigrants do not arrive as unemployed workers, but instead as entrepreneurs waiting to start a small business. This means that the expansion of employment may occur at the same pace as the increase in workers. In this case the economy moves directly along the profit curve from *A* to *C*, without dipping down to point B on the way.

DISCUSS 16.7: THE ECONOMIC EFFECTS OF IMMIGRATION

The economic effects of immigration are widely debated among the public. Watch this analysis of the economic effect of migration in Swindon, a town in Britain.

1. Summarise the evidence on migrants' skills suggested in the video.
2. Use the model explained in this section to show what may happen to wages and employment after an influx of migrant workers.
3. What is the evidence on the effect of immigration on wages in Britain reported in the video? Compare this with your prediction from question 2. Try to modify the model to come up with an explanation of this evidence.

16.8 GLOBALISATION AND ANTI-GLOBALISATION

As the 19th century examples of European agricultural protection and New World immigration restrictions show, globalisation can undermine itself. It produces winners and losers. By allowing countries to specialise in the production of goods for which they have a comparative advantage, globalisation of trade in goods and services can expand the consumption possibilities of all nations. But the freer movement of capital around the world in search of profit-making opportunities also allows businesses to seek countries with lax environmental regulation and low taxation. This, in turn, puts pressure on governments to adopt policies that fail to address questions of environmental quality and economic justice. The freer movement of goods and capital, as we saw in Units 13 to 15, also limits the effectiveness of policies to stabilise aggregate demand and employment. The movement of labour from one country to another creates gains for some, but threatens losses for others. If the losers are ignored globalisation may turn out to be politically unsustainable.

You may have noticed from our examples above that the distributional impact of immigration in a rich, capital-abundant country is the same as the distributional impact of trade. When Germany imports electronics, this lowers wages and increases profits. When China imports machine tools, this has the opposite effect on wages and profits in China. When Germany imports people, this also tends to lower wages and increase profits in the short run.

Thus importing goods of a type that takes a lot of people to produce (labour-intensive goods) is not very different economically from importing the people themselves. However, voters in the relatively rich countries seem to be much more hostile to immigration than to trade. Americans, for example, seem happy to purchase clothing made in Bangladesh at bargain prices, but are not pleased at the arrival of citizens from that country seeking work.

The discrepancy may be because of religious, cultural or ethnic prejudice, or because of a fear that valued local culture and social practices may be undermined. It may also be because immigrants consume public services paid for by taxation, while imported goods do not. Of course, immigrants also pay taxes and typically make a net positive contribution to the public finances, which imports do not (the public debate about immigration often ignores this fact).

Anti-globalisation sentiment is not just due to concerns about income distribution, or fears regarding immigration. Some worry that globalisation is placing too much power in the hands of large corporations, and placing democracy under threat.

These political concerns have been analysed by Dani Rodrik, an economist, who developed what he calls the *fundamental political trilemma of the world economy*. His trilemma refers to three things, all of which taken on their own seem good, but which cannot all occur at the same time. Rodrik's trilemma is really just another trade-off, like that between low inflation and low unemployment (it's hard to have both), or more free time and higher grades (more time at the beach is not the way to get the top grade); except that Rodrik's trade-off is in three dimensions.

He defines the three dimensions as:

1. *Hyperglobalisation*: A world in which there are virtually no political or cultural barriers to the location of goods and investment. They can be wherever in the world they, or their owners, choose.
2. *Democracy within nation states*: This means (as we said in Unit 1) that the government respects both individual liberty and political equality.
3. *National sovereignty*: Each national government can pursue policies that it chooses without any significant limits or assistance in enforcement by global institutions: restrictions on carbon emissions, the treatment of employees, the nature of intellectual property rights, or the payment of taxes for example.

According to Rodrik, hyperglobalisation means that countries have to compete with each other for investment, which will constantly seek locations in which labour and the environment are less regulated or are more competitive in other ways. This makes it difficult for national governments to adopt regulatory standards or other policies, or raise taxes on mobile capital or highly paid workers, even when citizens want it. For example, voters everywhere probably want to see tighter regulation on financial institutions, but governments worry about losing the financial sector to foreign competitors. In short, the threat of multinational companies and highly paid workers to leave trumps the voters' preferences.

Figure 16.21 illustrates the three possible outcomes of Rodrik's political trilemma.

First, governments may choose to limit globalisation so that it does not take its hyper form, in order to retain some freedom to set policies which correspond to preferences of their citizens. This prioritises national governments and national sovereignty. The rapid growth of Japan, South Korea and Taiwan—the Asian miracle—depended on well-designed use of protection from import competition, combined with incentives for exporting manufactures after 1945. Their achievements could not have been accomplished in the context of hyperglobalisation. But note that there were other countries, such as those in Latin America during the same period, where poorly designed policies of protection prevented them from achieving rapid growth.

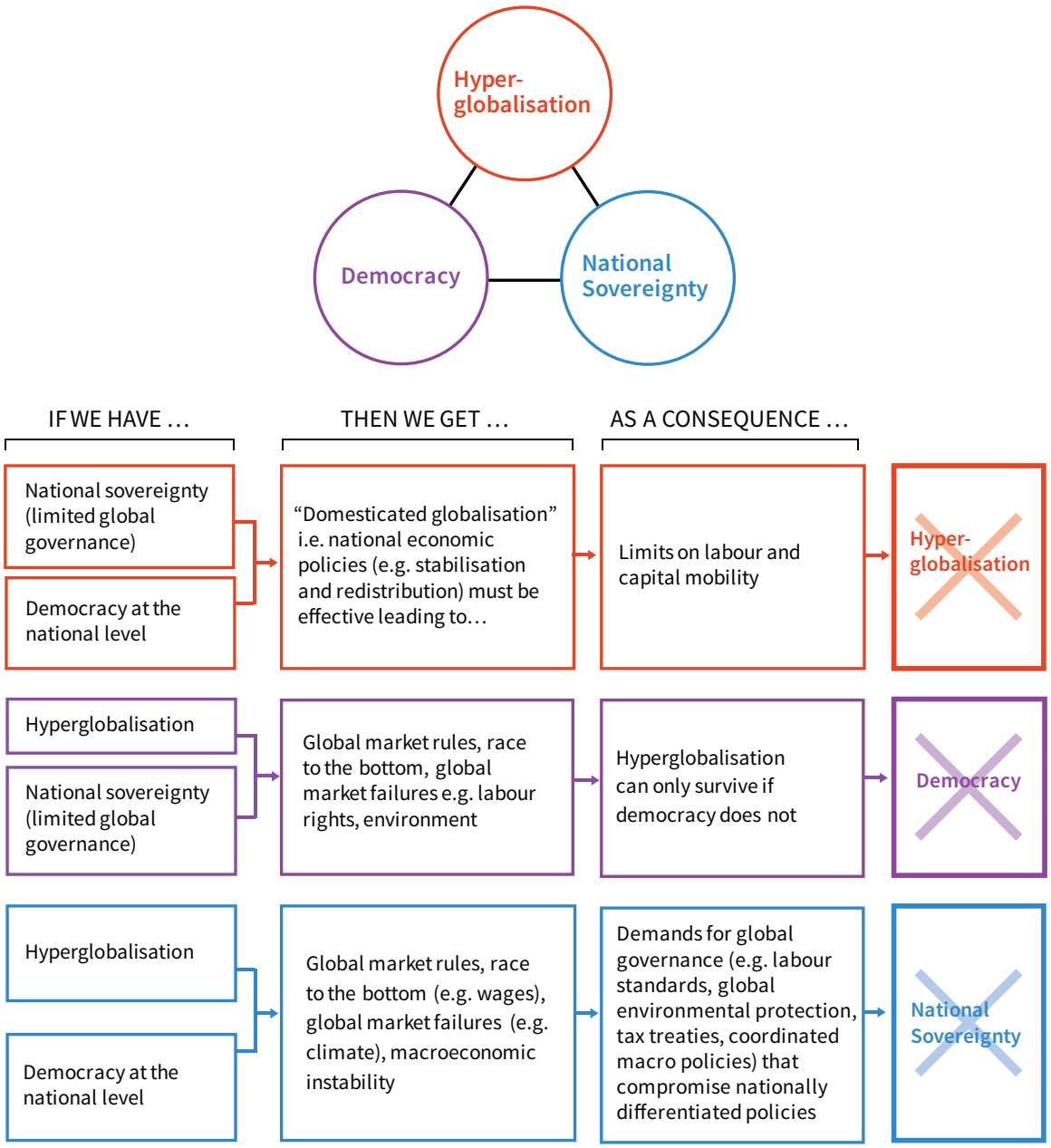


Figure 16.21 *Rodrik's political trilemma.*

Source: *Adapted from Rodrik, Dani. 2012. The Globalization Paradox: Democracy and the Future of the World Economy. United States: W. W. Norton & Company.*

For countries that embrace hyperglobalisation, according to Rodrik, there are two possible futures. Neither is very promising:

- *Governments choose hyperglobalisation and the sovereignty of the nation*: They will oppose supranational governance, such as any restrictions of the movement of capital globally, or international environmental agreements seeking to limit climate change. In the long run this could result in competitive deregulation and tax reductions, as well as increasing inequality between globalisation's winners and losers within countries. Many have labelled this a *race to the bottom* as public policy increasingly fails to address market failures, such as the environment, or

questions of economic fairness. The demands of hyperglobalisation may limit the use of macroeconomic stabilisation policy. This is a policy that the US has favoured in recent decades.

Note that others might call this a *golden straitjacket*, and welcome the restrictions that globalisation will impose on democratic governments who may wish to redistribute income from winners in the economic game to losers.

The outcome is likely to be both of these things, and it is not likely to be popular among the vast majority of citizens. To perpetuate an unpopular set of policies, those who advocate national sovereignty plus hyperglobalisation may eventually try to find ways of limiting democracy.

- *Governments may regulate markets at a regional or world level*: This mitigates the race-to-the-bottom downside of globalisation. An example is the political integration of Europe over the last few decades. It happened, in part, so that governments could obtain the benefits of free trade, plus the free movement of capital and labour, while retaining some ability at the regional level to regulate profitmaking in the interests of fairness and economic stability.

  The obvious problem is how to make sure that this regional or global governance is democratic as well as technocratic, to allow voters to change the system if they don't like it. Other supranational governance initiatives include world agreements on climate change, and efforts by the International Labour Organisation to require that all nations meet at least minimal standards for the treatment of labour (eliminating child labour or the physical coercion of employees, for example).

Watch as Rodrik explains that increasing globalisation implies that nations must "give up some sovereignty or some democracy".

DISCUSS 16.8: RODRIK'S TRILEMMA

Watch Dani Rodrik's video.

1. In your own words, explain what is meant by the globalisation trilemma.
2. Describe the main winners and losers from globalisation in a country.
3. How might governments try to overcome some of the downsides of globalisation?

16.9 TRADE AND GROWTH

What are the best policies for governments to adopt if they seek to promote long-run growth in living standards? Some argue that it is a choice between two policy extremes:

- Seal the national borders and withdraw from the world economy!
- Let trade, immigration, and investment across national boundaries be determined entirely by private individuals!

Few (if any) economists advocate either policy. The question is how to exploit the contributions of the global economy to a nation's growth, while minimising the ways in which integration into the global economy may retard growth. Among the growth-enhancing aspects of greater global economic integration are:

- *Competition*: Limiting the impediments to trade in goods and services among nations increases the degree of competition faced by firms in the local economy. This means that firms that fail to adopt new technologies and other cost-cutting methods are more likely to fail and to be replaced by more dynamic firms. The result will be an acceleration of the rate of technological progress.
- *Economies of scale*: A firm that can export to the world market has the opportunity (if it can meet the competition) of selling far more than it could were it restricted to the domestic market. This allows lower-cost production, which benefits home economy buyers, employees and owners of these successful firms, as well as external buyers.

Ways that greater integration into the global economy might retard growth include:

- *Learning by doing in infant industries*: In addition to economies of scale, another factor contributing to cost reductions is termed *learning by doing*. Even if the firm never achieves large-scale production, costs of production typically fall over time. Tariffs protecting *infant industries* can give firms the time and possibly the scale of operation necessary to become competitive.

INFANT INDUSTRY

A relatively new industrial sector in a country that has relatively high costs, because:

- Its recent establishment means that it has few benefits from *learning by doing*
- Its small size deprives it of *economies of scale*
- A lack of similar firms means that it does not benefit from *economies of agglomeration*

Temporary *tariff* protection of an infant industry may increase productivity in an economy in the long run.

- *Disadvantageous specialisation*: For reasons of history, some countries may specialise in sectors where there is a lot of potential for innovation, whereas others specialise in sectors with little such potential. Many Latin American countries, for example, slowed growth by specialising in low-innovation sectors such as natural resource extraction. Developing new specialisations may require direct government intervention, including infant industry protection.

It is clear from Figure 16.22 that during the second period of globalisation workers in some countries—China and South Korea for example—have seen rapid increases in their income levels. But the same figure also makes it clear that in other countries, such as Mexico and Sri Lanka, workers have seen few benefits from the increasingly integrated world economy.

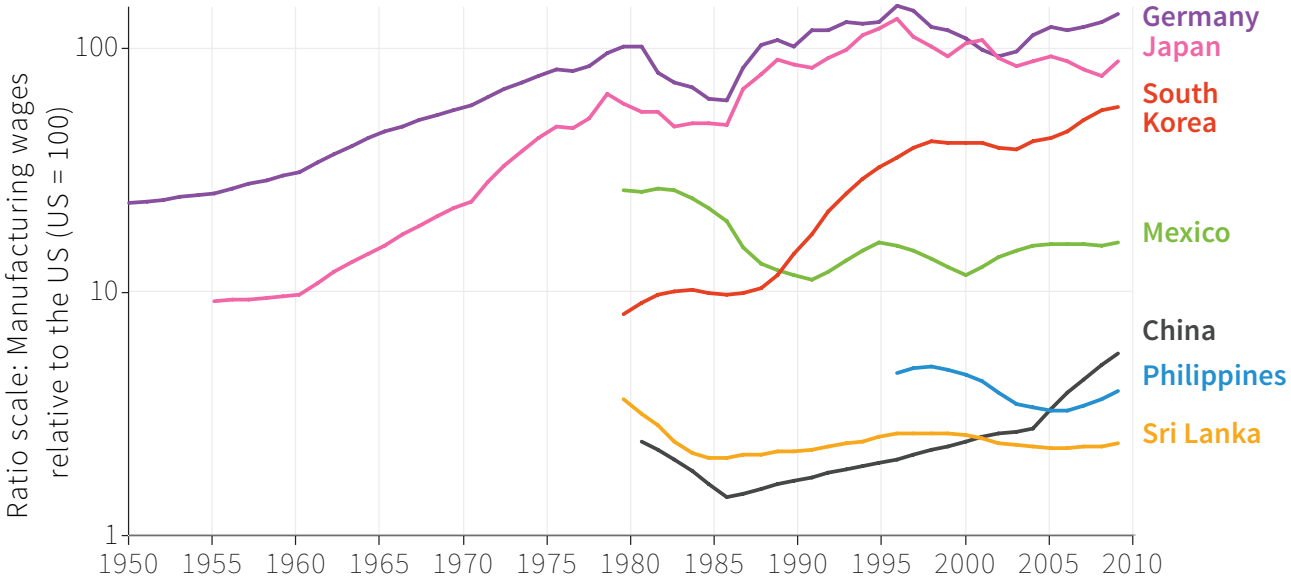


Figure 16.22 *Catching up and stagnating: Manufacturing wages relative to the US (1950-2009).*

Source: *Glyn, Andrew. 2006. Capitalism Unleashed: Finance, Globalization, and Welfare. Oxford: Oxford University Press. National Bureau of Statistics of China. 2015. 'Annual Data.' Bank of England. US Bureau of Labor Statistics. 2015. 'International Labor Comparisons.' Note: Annual US BLS data for Mexico, the Philippines and Sri Lanka has been smoothed using a backward-looking five-year moving average.*

There has not been a unique route to economic success during the past 150 years. For example:

- *Early protectionism in Germany and the US*: These countries developed modern manufacturing sectors behind high tariff barriers that sheltered them from British competition. In the late 19th century the correlation between tariffs and economic growth across relatively rich countries was positive. In particular, higher manufacturing tariffs were associated with higher growth. During the interwar period, tariffs were also positively correlated with growth.

- *Two directions after 1945*: On the one hand, countries in east Asia that encouraged their firms to compete in international markets grew faster than Latin American countries that were more closed to international trade. On the other hand, after those Latin American countries reduced their tariffs in the early 1990s, their subsequent economic growth rates were lower than the more closed period 1945 to 1980.
- *Scandinavian prosperity through openness*: These countries have been very open to trade for more than 100 years and have prospered. So as to mitigate the fluctuations in household income associated with changes in international prices, they also have very high tax rates to support generous social insurance and subsidies for retraining.
- *Picking national winners*: Many east Asian governments have promoted trade while influencing its pattern by favouring certain industries, or even certain firms, and by directing firms to compete in export markets whilst providing some protection from import competition.

If there is a lesson from this, it is that success does not depend on whether a country is more or less integrated into the world economic system—more or fewer exports and imports, for example, or a greater amount of international investment by its firms—but instead it depends on how economic integration is managed by policy to promote growth.

DISCUSS 16.9: THE EFFECT OF TRADE ON GROWTH

The empirical evidence on how trade affects growth is mixed.

1. Suppose you are a consultant with the World Trade Organisation and are asked to design an empirical study to find the effect of a country's openness to trade on growth. How would you approach this exercise?
2. Justify your choices about how you would measure openness to trade. (Would you use tariffs, export ratios or other indices of openness? Why?)
3. Explain the problems you face in designing a convincing study.

16.10 CONCLUSION

In an essay titled *National Self-Sufficiency,* published in 1933, John Maynard Keynes warned of the consequences of globalisation before the word even existed:

"We each have our own fancy. Not believing that we are saved already, we each should like to have a try at working out our own salvation. We do not wish, therefore, to be at the mercy of world forces working out, or trying to work out some uniform equilibrium according to the ideal principles, if they can be called such, of laissez-faire capitalism... We wish for the time at least... to be our own masters and to be as free as we can... to make our own favourite experiments towards the ideal social republic of the future."
— John Maynard Keynes, *National Self-Sufficiency* (1933)

It became conventional wisdom that global integration would eventually make the idea of national economic sovereignty impractical. Almost half a century after Keynes wished for time "to be our own masters", Charles Kindleberger, an international trade economist, argued that:

"The nation state is just about through as an economic unit... It is too easy to get about. Two-hundred-thousand-ton tankers... airbuses and the like will not permit the sovereign independence of the nation-state in economic affairs."
— Charles Kindleberger, *American Business Abroad: Six Lectures on Direct Investment* (1969)

But were Keynes and Kindleberger right to think that globalisation would reduce all nations to minor players in the world economy?

We have seen that the world's economies are now part of an integrated global system. Major companies consider the entire world when deciding where to produce and where to sell their goods and services. Investors, likewise, choose where to hold their assets, whether financial or

CONCEPTS INTRODUCED IN UNIT 16

Before you move on, review these definitions:

- *Globalisation* and *Hyperglobalisation*
- *Specialisation*
- *Comparative advantage*
- *Price gap, Trade costs*
- *Arbitrage*
- *Globalisation I and II*
- *Tariff*
- *Current account* (CA), *CA deficit, CA surplus, Net capital flows*
- *Balance of payments* accounts
- *International capital flows*
- *Gains from trade*
- *Foreign direct investment*
- *Foreign portfolio investment*
- *Economies of agglomeration*
- *Learning by doing*
- *Infant industries*

real, on the basis of calculations of expected returns in all the regions of the world. But we have also seen that labour has for the most part not been globalised, and for political, cultural and language reasons remains largely national. National borders remain an essential fact of the global economy. National governments are major actors in affecting the course of their own and other economies.

As Keynes and Kindleberger anticipated, globalisation has brought about important changes. In the 18th century, at the birth of economics as a discipline, goods were traded across national boundaries, and investments were made in far-flung parts of the world; but for the most part the nation and its economy had the same boundaries.

The world today looks quite different. Trading of goods and services and investment are now integrated into the world economic system. In this sense we can think of the world economy (with few exceptions such as communist North Korea) as a single capitalist system. But the labour markets of the world are still sharply separated, and the policies and institutions adopted governing the markets in labour, capital and goods are still strongly shaped by national governments.

Key points in Unit 16

Declining price gaps

We see these for most goods and services. They are an indication of global integration of economies.

Comparative advantage

If countries specialise in the production of goods and services in which they have a comparative advantage, trade expands the consumption possibility frontier, an effect similar to that of technological progress.

Tariffs may impede specialisation

Tariffs and other policies often impede this specialisation process, resulting in forgone mutual gains.

Tariffs may protect infant industries

If infant industries are temporarily subsidised or protected by tariffs, then firms can reduce costs over time as they benefit from learning by doing, economies of scale and economies of agglomeration.

A country's comparative advantage

This depends on not only on the abundance of capital, labour, land and other resources but also institutions, culture, and public policy.

Conflicts over the distribution of gains from trade

Both within and between countries conflicts arise over the distribution of mutual gains made possible by specialisation and trade.

Rodrik's trilemma

Countries may not be able simultaneously to achieve complete globalisation, democracy and national sovereignty.

Globalisation in context

Freer movement of goods and services and of capital may promote more rapid economic growth under some conditions but, under other conditions, may retard growth.

16.11 READ MORE

Bibliography

1. Carter, Susan B., Michael R. Haines, Richard Sutch, and Scott Sigmund Gartner, eds. 2006. *Historical Statistics of the United States: Earliest Times to the Present*. New York: Cambridge University Press.
2. Glyn, Andrew. 2006. *Capitalism Unleashed: Finance, Globalization, and Welfare*. Oxford: Oxford University Press.
3. Helpman, Elhanan. 1999. 'The Structure of Foreign Trade.' *Journal of Economic Perspectives* 13 (2): 121–44.
4. International Monetary Fund. 2014. *World Economic Outlook April: Recovery Strengthens, Remains Uneven*. Washington, DC: IMF.
5. International Monetary Fund. 2014. 'World Economic Outlook Database: October 2014.'
6. Jacks, David S., Christopher M. Meissner, and Dennis Novy. 2011. 'Trade Booms, Trade Busts, and Trade Costs.' *Journal of International Economics* 83 (2): 185–201.
7. Keynes, John Maynard. 1933. 'National Self-Sufficiency.' *The Yale Review* 22 (4): 755–69.
8. Kindleberger, Charles. 1969. *American Business Abroad: Six Lectures on Direct Investment*. United States: Yale University Press.
9. Krugman, Paul. 2009. 'The Increasing Returns Revolution in Trade and Geography.' In *The Nobel Prizes 2008*, edited by Karl Grandin. Stockholm: The Nobel Foundation.

10. Krugman, Paul R. 1987. 'Is Free Trade Passé?' *Journal of Economic Perspectives* 1 (2): 131–44.
11. Lane, Philip R., and Gian-Maria Milesi-Ferretti. 2007. 'Europe and Global Imbalances.' *IMF Working Papers* 07 (144).
12. Maddison, Angus. 1995. *Monitoring the World Economy, 1820-1992*. Washington, DC: Development Centre of the Organisation for Economic Co-operation and Development.
13. Maddison, Angus. 2001. *The World Economy: A Millennial Perspective (Development Centre Studies)*. Paris: Organisation for Economic Co-operation and Development.
14. National Bureau of Statistics of China. 2015. 'Annual Data.'
15. O'Rourke, Kevin, and Jeffrey G. Williamson. 1994. 'Late Nineteenth-Century Anglo-American Factor-Price Convergence: Were Heckscher and Ohlin Right?' *The Journal of Economic History* 54 (04): 892–916.
16. O'Rourke, Kevin H., and Jeffrey G. Williamson. 2005. 'From Malthus to Ohlin: Trade, Industrialisation and Distribution since 1500.' *Journal of Economic Growth* 10 (1): 5–34.
17. Obstfeld, Maurice, and Alan M. Taylor. 2005. *Global Capital Markets: Integration, Crisis, and Growth (Japan-US Center UFJ Bank Monographs on International Financial Markets)*. Cambridge: Cambridge University Press.
18. Ricardo, David. 1815. *An Essay on Profits*. London: John Murray.
19. Ricardo, David. 1817. *The Principles of Political Economy and Taxation*. London: John Murray.
20. Rodrik, Dani. 2012. *The Globalization Paradox: Democracy and the Future of the World Economy*. United States: W. W. Norton & Company.
21. The World Bank. 2011. 'Data on Trade and Import Barriers.'
22. Trefler, Daniel. 1995. 'The Case of the Missing Trade and Other Mysteries.' *The American Economic Review* 85 (5): 1029–46.
23. US Bureau of Labor Statistics. 2015. 'International Labor Comparisons.'
24. United Nations Conference on Trade and Development. 2014. 'Bilateral FDI Statistics.'
25. World Trade Organization. 2013. *World Trade Report*. Geneva: WTO.

Beta February 2016 version

THE GREAT DEPRESSION, THE GOLDEN AGE OF CAPITALISM AND THE GLOBAL FINANCIAL CRISIS

SINCE THE END OF THE FIRST WORLD WAR, THREE PERIODS OF DOWNTURN AND INSTABILITY HAVE PUNCTUATED THE ECONOMIC HISTORY OF THE ADVANCED CAPITALIST ECONOMIES, INTERRUPTING LONG PERIODS OF RELATIVELY STEADY GROWTH IN LIVING STANDARDS. ECONOMISTS HAVE LEARNED DIFFERENT LESSONS FROM EACH OF THESE CRISES

- There have been three distinctive economic epochs in the hundred years following the first world war—the roaring twenties and the Great Depression; the golden age of capitalism and stagflation; and the great moderation and subsequent financial crisis of 2008
- The end of each of these epochs—the stock market crash of 1929; the decline in profits and investment in the late 1960s and early 1970s culminating in the oil shock of 1973; and the financial crisis of 2008—was a sign that institutions that had governed the economy to that point had failed
- The new institutions marking the golden age of capitalism—increased trade union strength and government spending on social insurance—addressed the aggregate demand problems highlighted by the Great Depression and were associated with rapid productivity growth, investment and falling inequality
- Nevertheless, the golden age ended with a crisis of profitability, investment and productivity followed by stagflation
- The policies adopted in response to the end of the golden age restored high profits and low inflation, but did not restore the investment and productivity growth of the previous epoch—and made economies vulnerable to debt-fuelled financial booms. One of these booms precipitated a global financial crisis in 2008

See www.core-econ.org for the full interactive version of *The Economy* by The CORE Project.
Guide yourself through key concepts with clickable figures, test your understanding with multiple choice questions, look up key terms in the glossary, read full mathematical derivations in the Leibniz supplements, watch economists explain their work in Economists in Action – and much more.

Institute for New Economic Thinking
UCL SciencesPo Azim Premji University
FRIENDS PROVIDENT Foundation

Before dawn on Saturday, 7 February 2009, 3,582 firefighters began deploying across the Australian state of Victoria. It would be the day remembered by Australians as Black Saturday: the day that bushfires devastated 400,000 hectares, destroyed 2,029 homes, and took 173 lives.

But when the fire brigades suited up that morning, there had not been any reports of fire. What had mobilised every firefighter in Victoria was the McArthur Forest Fire Danger Index (FFDI), which the previous day exceeded what, until then, was its calibrated maximum of 100—a level that had been reached only during the bushfires of January 1939. When the FFDI exceeds 50, it indicates "extreme" danger. A value above 100 is "catastrophic" danger. On 6 February 2009 it had hit 160.

Later there would be accusations, trials and even a Royal Commission to determine who or what had caused Australia's worst natural disaster. There were many possible causes: lightning strikes, sparks from farm machinery, faulty power lines, even arson.

A single spark or a lightning strike did not cause Black Saturday. Every day sparks ignite small bush fires, and on that day alone the Royal Commission reported 316 separate grass, scrub or forest fires. This was not a calamity because of any one of these local fires, but because conditions transformed routine, easily contained bushfires into an unprecedented disaster.

Small causes are sometimes magnified into large effects. Avalanches are another natural example. In electricity grids a failure of one link in the network overloads other links, leading to a cascade of failures and a blackout.

This small-causes-with-big-consequences process is found in economics too, for example in the Great Depression of the 1930s and the *global financial crisis* of 2008.

Although recessions are characteristic of capitalist economies, as we have seen, they rarely turn into episodes of persistent contraction. This is because of a combination of the economy's self-correcting properties and successful intervention by policymakers. Specifically:

- Households take preventative measures that dampen rather than amplify shocks (Unit 12)
- Governments create automatic stabilisers (Unit 13)
- Governments and central banks take actions to produce negative rather than *positive feedbacks* when shocks occur (Units 13, 14)

But, like Black Saturday, occasionally a major economic calamity occurs. These calamities raise three sets of questions about how economic crises mirror these natural disasters:

- In economics, what is the counterpart to the dry undergrowth, the small spark, and the positive feedback processes that caused the fire to spread? What creates the raw material for an economic "Black Saturday"?

- Do we wait for the fire burn out, or can we put it out? If so, how?
- How can the lessons of an economic crisis be used to reduce the chance it will happen again? Can a long period without a disaster lead to complacency?

In this unit we look at three crises that have punctuated the last century of unprecedented growth in living standards in the rich countries of the world—the Great Depression of the 1930s, the end of the golden age of capitalism in the 1970s, and the global financial crisis of 2008.

The global financial crisis in 2008 took households, firms and governments around the world by surprise. An apparently small problem in an obscure part of the housing market in the US caused house prices to plummet, leading to a cascade of unpaid debts around the world, and a collapse in global industrial production and world trade.

To economists and historians, the events of 2008 looked scarily like what had happened at the beginning of the Great Depression in 1929. For the first time they found themselves fretting about the level of the little-known *Baltic Dry Index* (you can track its current level here), a measure of shipping prices for commodities like iron, coal and grain. When world trade is booming, demand for these commodities is high. But the supply of freight capacity is inelastic, so shipping prices rise and the Index goes up. In May 2008 the Baltic Dry Index reached its highest level since it was first published in 1985. But the reverse is also true: by December many more people were checking the Index, because it had fallen 94%. The fall told them that, thousands of miles from the boarded-up houses of bankrupt former homeowners in Arizona and California where the crisis had begun, giant $100m freighters were stuck in port because there was no trade for them to carry.

In 2008 economists remembered the lessons of the Great Depression. They encouraged policymakers globally to adopt a coordinated set of actions to halt the collapse in aggregate demand, and to keep the banking system functioning.

But economists also share some of the responsibility for the policies that made this crisis more likely. For 30 years unregulated financial and other markets had been stable. Some economists incorrectly assumed that they were immune to instability. So the events of 2008 also show how a failure to learn from history helps to create the next crisis.

How did a small problem in the US housing market send the global economy to the brink of a catastrophe?

- *The dry undergrowth:* In Unit 16, Figure 16.9 charted the growth in the globalisation of international capital markets by looking at the amount of foreign assets owned by domestic residents. At the same time, the globalisation of banking was occurring. Some of the unregulated expansion of lending by global banks ended up financing mortgage loans to so-called *subprime* borrowers in the US.

- *The spark:* Falling real estate prices meant that banks with very high leverage, and therefore with thin cushions of net worth (equity), in the US, France, Germany, the UK and elsewhere quickly because insolvent.
- *The positive feedback mechanism:* Fear was transmitted around the world and customers cancelled orders. Aggregate demand fell sharply. The high degree of interconnection among global banks and the possibility of massive transactions in a matter of seconds made excessive leverage an increasingly dangerous source of instability.
- *The complacent policymakers:* With few exceptions most policymakers, and the economists whose advice they sought, still believed that the financial sector was able to regulate itself. The international central bank for central banks—the Bank for International Settlements in Basel—allowed banks great scope to choose their level of leverage. Banks could use their own models to calculate the riskiness of their assets. They could meet the international regulatory standards for leverage by understating the riskiness of their assets, and by parking these risky assets in what are called *shadow banks*, which they owned but which were outside the scope of banking regulations. All of this was entirely legal. Many economists continued to believe that economic instability was a thing of the past, right up to the onset of the crisis itself. It is as if Australian firefighters had watched the Forest Fire Danger Index hit 160, but did nothing because they didn't believe a fire was possible.

Some of those involved admitted afterwards that their belief in the stability of the economy had been wrong. For example, Alan Greenspan, who had been in charge of the US central bank between 1987 and 2006, admitted this error to a US government committee hearing.

As the financial crisis unfolded in the summer and autumn of 2008, economists in government, central banks and universities diagnosed a crisis of aggregate demand and bank failure. Many of the key policymakers in this crisis were economists who had studied the Great Depression.

The lessons they had learned from the Great Depression in the US—cut interest rates, provide liquidity to banks and run fiscal deficits—were applied. In November 2008, ahead of the G20 summit in Washington, British Prime Minister Gordon Brown told reporters:

"We need to agree on the importance of coordination of fiscal and monetary policies. There is a need for urgency. By acting now we can stimulate growth in all our economies. The cost of inaction will be far greater than the cost of any action."

HOW ECONOMISTS LEARN FROM FACTS

"I MADE A MISTAKE"

On 23 October 2008, a few weeks after the collapse of the US investment bank Lehman Brothers, former US Federal Reserve chairman Alan Greenspan admitted that the accelerating financial crisis had shown him "a flaw" in his belief that free, competitive markets would ensure financial stability. In a hearing of the US House of Representatives Committee on Oversight and Government Reform, Greenspan was questioned by chair of the House Committee, Rep. Henry Waxman:

Waxman Well, where did you make a mistake then?

Greenspan I made a mistake in presuming that the self-interest of organisations, specifically banks and others, was best capable of protecting [the banks'] own shareholders and their equity in the firms… So the problem here is that something which looked to be a very solid edifice, and, indeed, a critical pillar to market competition and free markets, did break down. And I think that, as I said, shocked me. I still do not fully understand why it happened and, obviously, to the extent that I figure out where it happened and why, I will change my views. If the facts change, I will change.

Waxman You had a belief that *[quoting Greenspan]* "free, competitive markets are by far the unrivalled way to organise economies. We have tried regulation, none meaningfully worked.'' You have the authority to prevent irresponsible lending practices that led to the subprime mortgage crisis. You were advised to do so by many others. [Did you] make decisions that you wish you had not made?

Greenspan Yes, I found a flaw…

Waxman You found a flaw?

Greenspan I found a flaw in the model… that defines how the world works, so to speak.

Waxman In other words, you found that your view of the world was not right, it was not working.

Greenspan Precisely. That's precisely the reason I was shocked, because I had been going for 40 years or more with very considerable evidence that it was working exceptionally well.

A direct comparison between the first 10 months of the Great Depression and the 2008 financial crisis shows that the collapse of industrial production in the world economy was similar (compare January 1930 and January 2009 in Figure 17.1a). But lessons had been learned: in 2008, monetary and fiscal policy responses were much larger and more decisive than in 1930, as shown in Figures 17.1b and 17.1c.

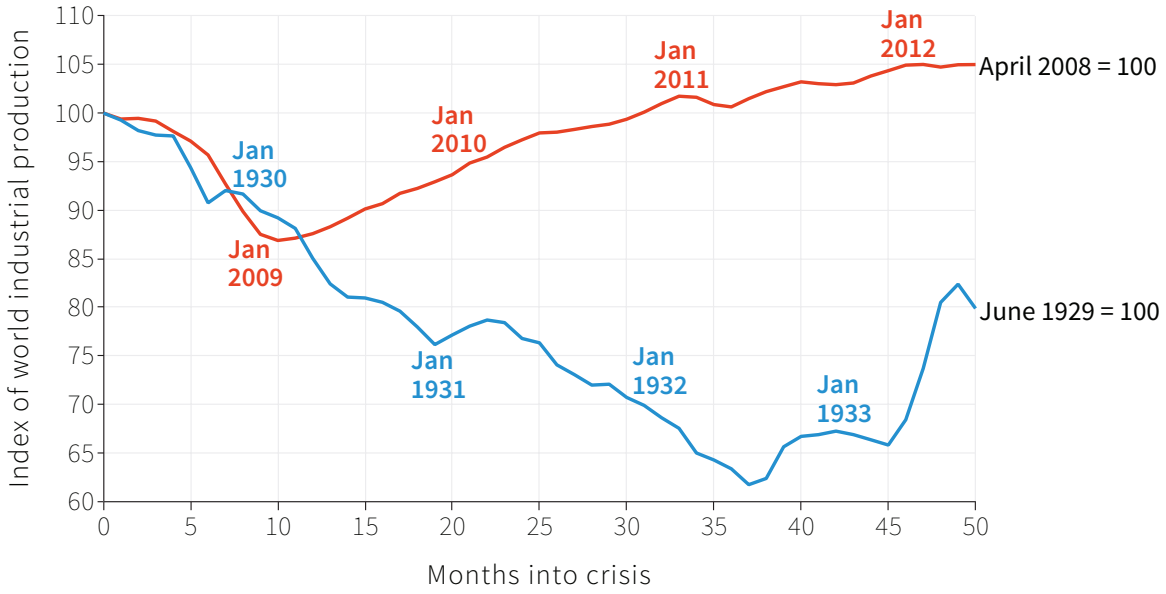


Figure 17.1a *The second Great Depression that did not happen: Comparing industrial production in the Great Depression and the global financial crisis.*

Source: *Almunia, Miguel, Agustín Bénétrix, Barry Eichengreen, Kevin H. O'Rourke, and Gisela Rua. 2010. 'From Great Depression to Great Credit Crisis: Similarities, Differences and Lessons.' Economic Policy 25 (62): 219–65. Updated using CPB Netherlands Bureau for Economic Policy Analysis. 2015. 'World Trade Monitor.'*

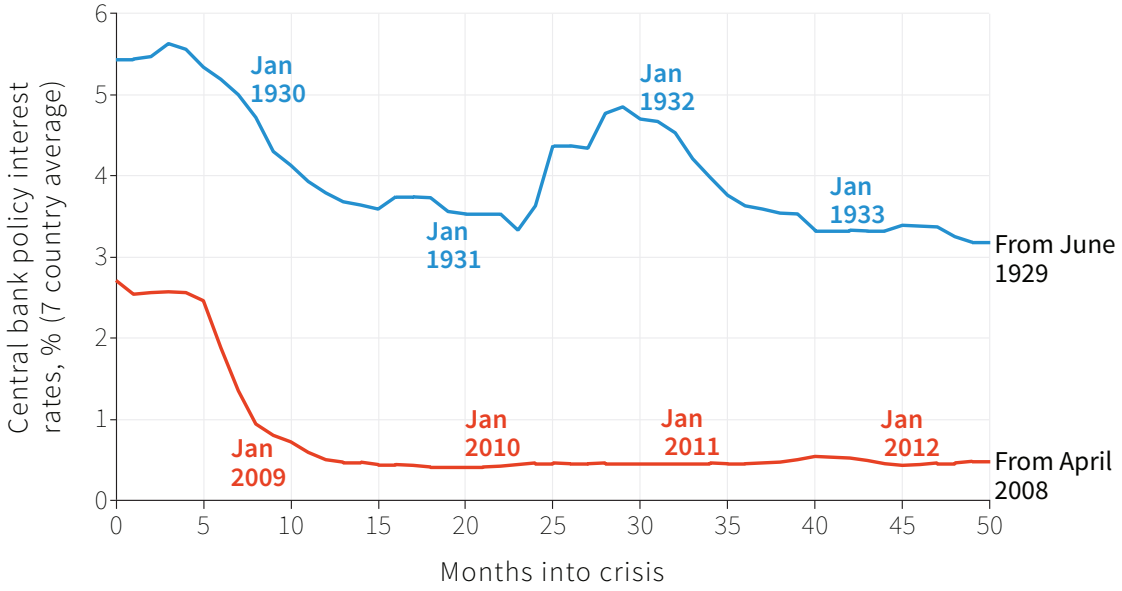


Figure 17.1b *The Great Depression and the global financial crisis: Monetary policy.*

Source: *As in Figure 17.1a updated using national central bank data.*

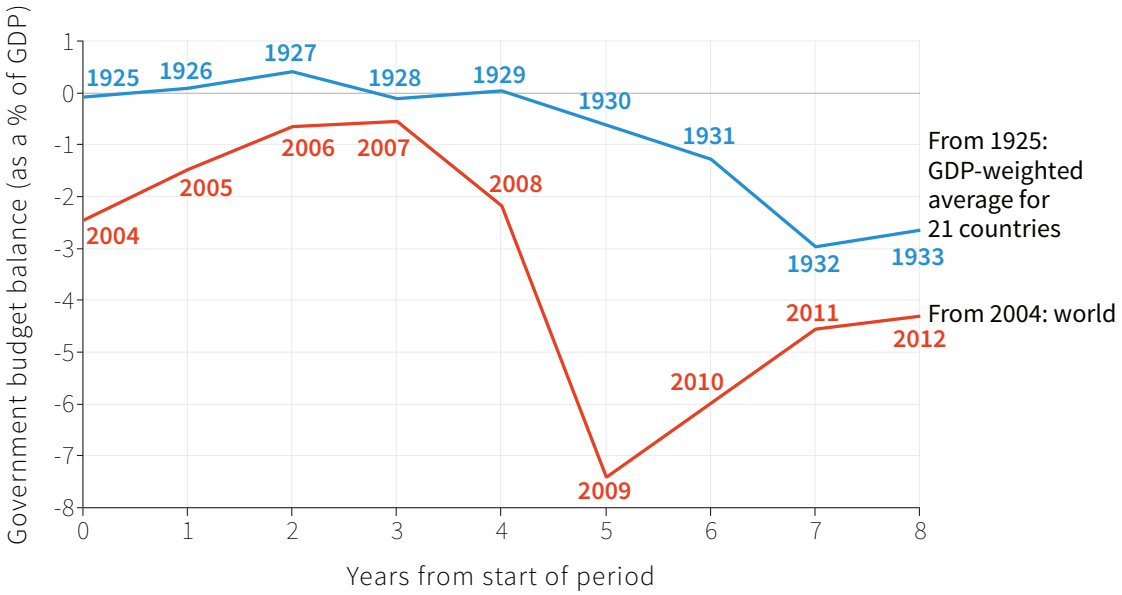


Figure 17.1c *The Great Depression and the global financial crisis: Fiscal policy.*

Source: *As in Figure 17.1a updated using International Monetary Fund. 2009. World Economic Outlook: January 2009; International Monetary Fund. 2013. 'IMF Fiscal Monitor April 2013: Fiscal Adjustment in an Uncertain World, April 2013.' April 16.*

17.1 THREE ECONOMIC EPOCHS

In the past 100 years the economies we often refer to as advanced (meaning, basically, "rich"), including the US, Western Europe, Australia, Canada and New Zealand, have seen average living standards measured by output per capita grow six-fold. Over the same period hours of work have fallen. This is a remarkable economic success, but it has not been a smooth ride.

The story of how rapid growth began was told in Units 1 and 2. In Figure 12.2 we contrasted the steady long-run growth rate from 1921 to 2011 with the fluctuations of the business cycle, which go from peak to peak every three to five years.

In this unit we will study three distinctive epochs. Each begins with a period of good years (the light shading in Figure 17.2 below), followed by a period of bad years (the dark shading).

- *1921 to 1941:* The crisis of the Great Depression is the defining feature of the first epoch, and opened the way for Keynes' concept of *aggregate demand* to become standard in economics teaching and policymaking.

- *1948 to 1979*: The golden age epoch stretched from the end of the second world war to 1979, and is named for the economic success of the 1950s and 1960s. The golden age was brought to an end in the 1970s by a crisis of profitability and productivity, and saw the emphasis in economics teaching and policymaking shift away from the role of aggregate demand toward *supply-side problems* such as productivity and decisions by firms to enter and exit markets.
- *1979 to 2013*: In the most recent epoch, the world was caught by surprise by the global financial crisis. The potential of a debt-fuelled boom to cause havoc was neglected during the preceding years of stable growth and seemingly successful macroeconomic management, which had been called the *great moderation*.

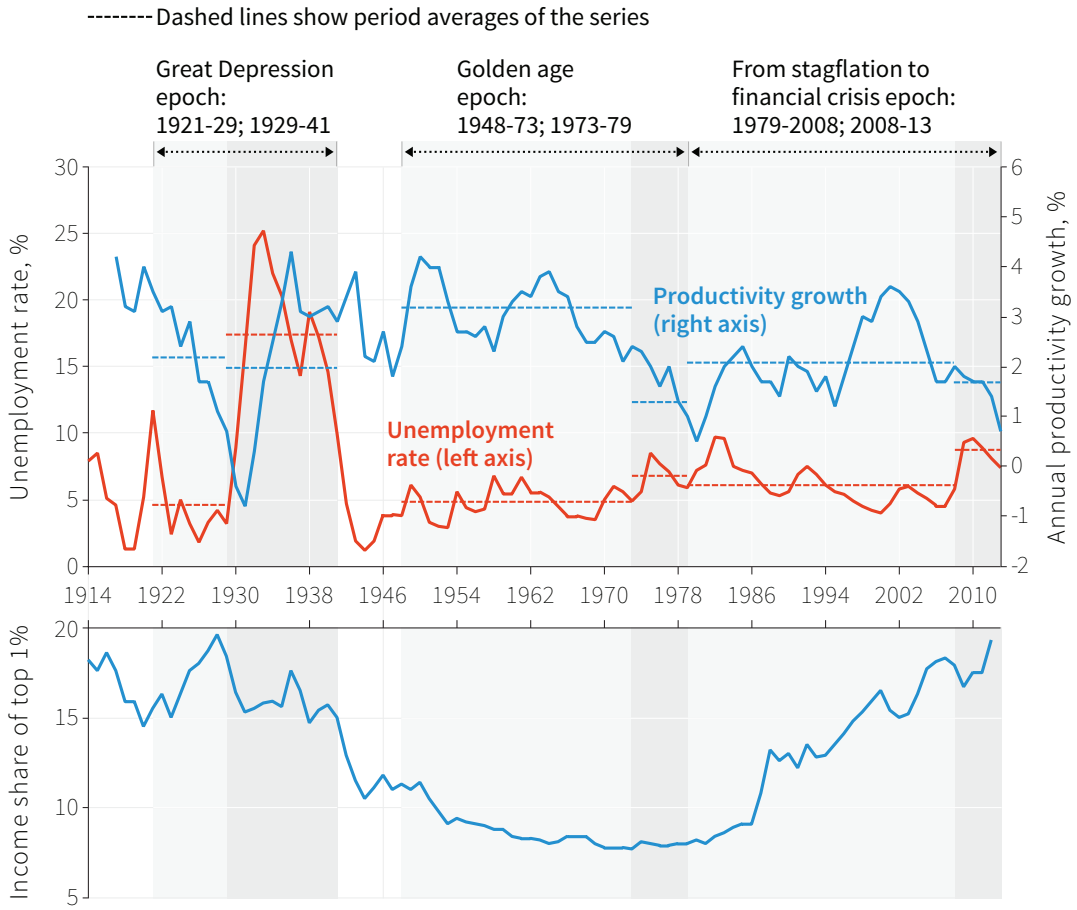


Figure 17.2 *Unemployment, productivity growth and inequality in the United States (1914-2013).*

Source: *United States Bureau of the Census. 2003. Historical Statistics of the United States: Colonial Times to 1970, Part 1. United States: United States Govt Printing Office; Alvaredo, Facundo, Anthony B Atkinson, Thomas Piketty, Emmanuel Saez, and Gabriel Zucman. 2016. 'The World Wealth and Income Database (WID).'; US Bureau of Labor Statistics; US Bureau of Economic Analysis.*

The term *crisis* is routinely applied to the first and the last of these episodes because they represented an unusual but recurrent cataclysmic divergence from the normal ups-and-downs of the economy. In the second epoch the end of the golden age, too, marked a sharp deviation from what had become normal. The three unhappy

surprises that ended the epochs are different in many respects, but they share a common feature: positive feedbacks magnified the effects of routine shocks that under other circumstances would have been dampened.

What does Figure 17.2 show?

- *Productivity growth:* A broad measure of economic performance is the growth of hourly productivity in the business sector. Productivity growth hit low points in the Great Depression, at the end of the golden age epoch in 1979 and in the wake of the financial crisis. The golden age got its name due to the extraordinary productivity growth until late in that epoch. The dashed blue lines show the average growth of productivity for each sub-period.
- *Unemployment:* High unemployment, shown in red, dominated the first epoch. The success of the golden age was marked by low unemployment as well as high productivity growth. The end of the golden age produced spikes in unemployment in the mid 1970s and early 1980s. In the third epoch, unemployment was lower at each successive business cycle trough until the financial crisis, when high and persistent unemployment re-emerged.
- *Inequality*: Figure 17.2 also presents data on inequality for the US: the income share of the top 1%. The richest 1% had nearly one-fifth of income in the late 1920s just before the Great Depression. Their share then steadily declined until a U-turn at the end of the golden age restored the income share of the very rich to 1920s levels.

We saw in earlier units that continuous technological progress has characterised capitalist economies, driven by the incentives to introduce new technology. Based on their expected profits after tax, entrepreneurs make investment decisions to get a step ahead of their competitors. Productivity growth reflects their collective decisions to invest in new machinery and equipment embodying improvements in technology. Figure 17.3 shows the growth rate of the capital stock and the rate of profit of firms in the non-financial corporate sector of the US economy before and after the payment of taxes on profits.

The data in Figure 17.3 illustrates that capital stock growth and firm profitability tend to rise and fall together. As we saw in Unit 13, investment is a function of expected post-tax profits and expectations will be influenced by what has happened to profitability in the recent past. Once a decision to invest is taken, there is a lag before the new capital stock is ordered and installed.

As profitability was restored following the collapse of the stock market in 1929 and the banking crises of 1929-31, investment recovered and the capital stock began to grow again. During the golden age, profitability and investment were both buoyant. A closer look at Figure 17.3 is revealing: investment depends on post-tax profitability and we can see that the gap between the pre-tax (red) and post-tax (green) rate of profit declined during the golden age. The lower panel shows the *effective tax rate* on corporate profits explicitly.

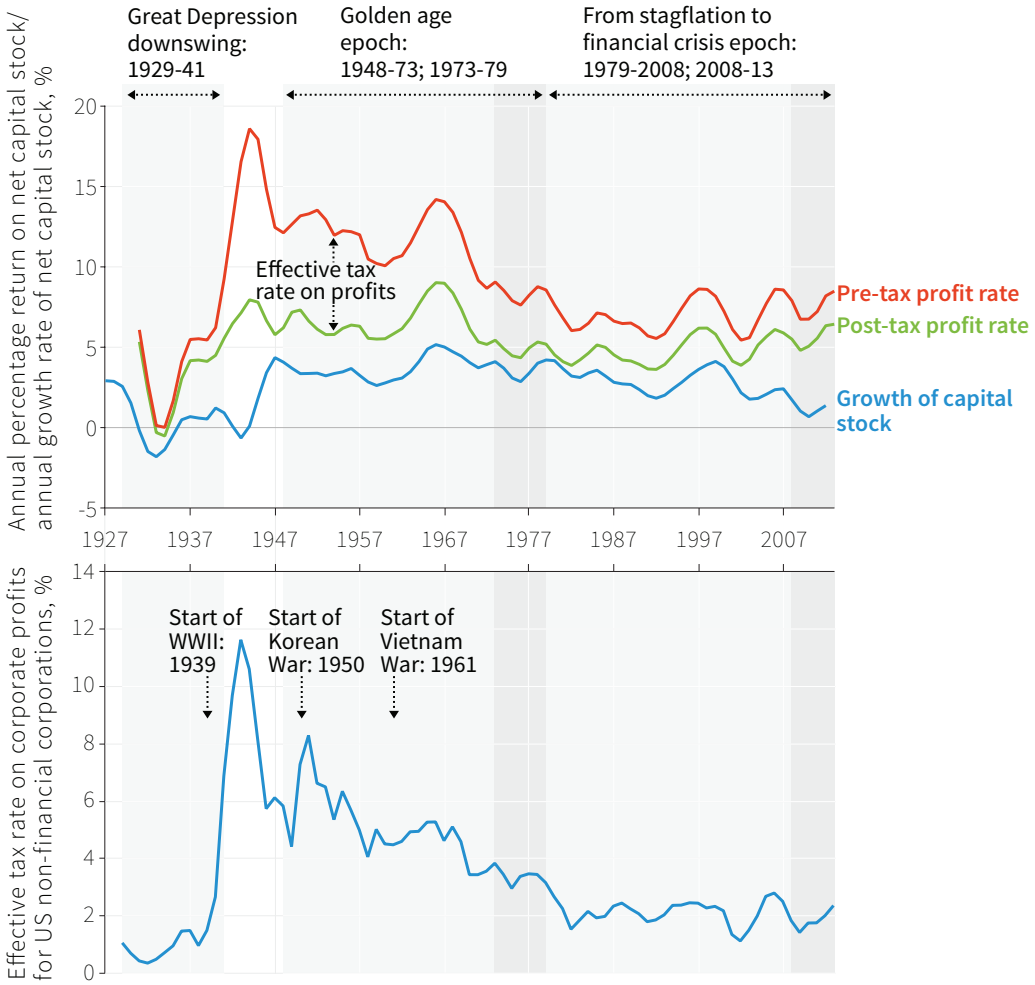


Figure 17.3 *Capital stock growth, profit rates and effective tax rate on profits for US non-financial corporations (1927-2013).*

Source: *US Bureau of Economic Analysis.*

Wars have to be financed and the tax on businesses increased during the second world war and the Korean war, and more slowly over the course of the Vietnam war. The effective tax rate on profits fell from 8% to 2% over the 30 years from the early 1950s. This helped to stabilise the post-tax rate of profit. In the late 1970s and early 1980s, taxes on profits were cut sharply; thereafter the pre-tax profit rate fluctuated without a trend. But in spite of the stabilisation of profitability in the third epoch, the growth rate of the capital stock fell.

On the eve of the financial crisis, Figures 17.2 and 17.3 show that the richest Americans were doing very well. But this did not stimulate investment, with the capital stock growing more slowly than at any time since the second world war. The onset of the financial crisis also coincided with a peak in debt (shown in Figure 17.4). Debt in financial firms and in households was at postwar highs (relative to the size of GDP). The swelling in the amount of debt was clearest for financial firms—but households also increased their debt ratio steadily through the 2000s.

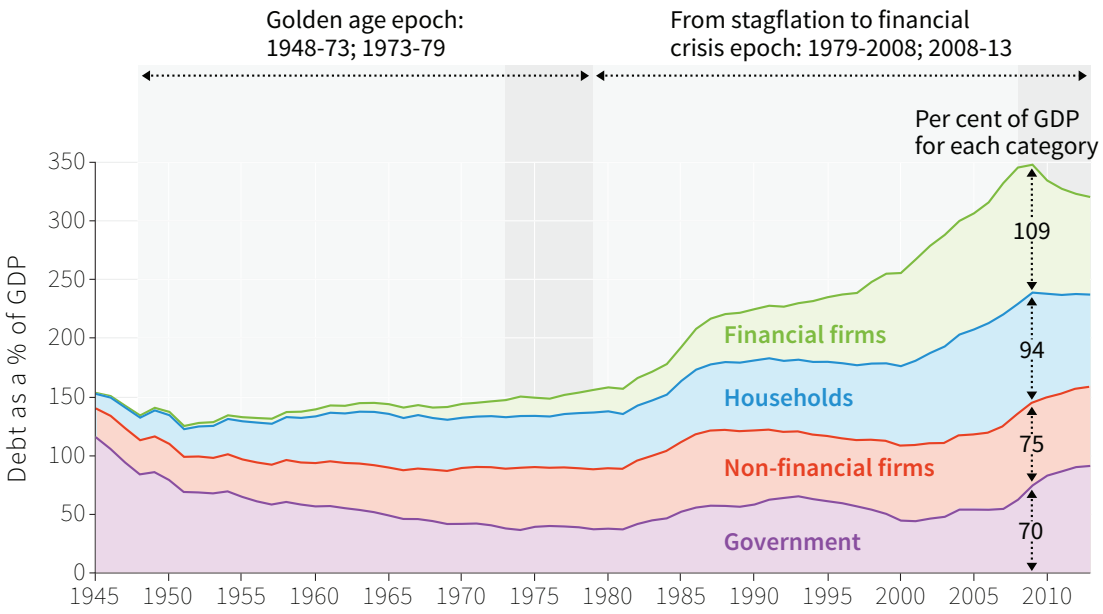


Figure 17.4 *Debt as a percentage of GDP in the United States: Households, non-financial firms, financial firms and the government (1945-2013).*

Source: *US Federal Reserve. 2015. 'Financial Accounts of the United States, Historical.' December 10; US Bureau of Economic Analysis.*

The three epochs of modern capitalism are very different, as Figures 17.5a and 17.5b show. We need to use the full range of tools of analysis we have developed in previous units to understand their dynamics, and how one epoch is related to another.

By 1921, the US had been the world productivity leader for a decade, and the world's largest economy for 50 years. The three epochs are more clearly defined in the US than in other countries, even other rich countries, although they had a profound influence on the economic history of the rest of the world. Its global leadership in technology and its global firms help explain rapid catch-up growth in Europe and Japan in the golden age. On either side of the golden age, the crises that began in the US in 1929 and 2008 became global crises too.

So why, apart from its productivity leadership, were these epochs centred on the US? Figure 17.5b summarises important differences between the US and other rich countries.

NAME OF PERIOD	DATES	IMPORTANT FEATURES OF THE US ECONOMY
1920s	1921-1929	• Low unemployment • High productivity growth • Rising inequality
GREAT DEPRESSION	1929-1941	• High unemployment • Falling prices • Unusually low growth rate of business capital stock • Falling inequality
GOLDEN AGE	1948-1973	• Low unemployment • Unusually high productivity growth • Unusually high growth rate of capital stock • Falling effective tax rate on corporate profits • Falling inequality
STAGFLATION	1973-1979	• High unemployment and inflation • Low productivity growth • Lower profits
1980s & THE GREAT MODERATION	1979-2008	• Low unemployment and inflation • Falling growth rate of business capital stock • Sharply rising inequality • Rising indebtedness of households and banks
FINANCIAL CRISIS	2008-2013	• High unemployment • Low inflation • Rising inequality

Figure 17.5a *The performance of the US economy over a century.*

GREAT DEPRESSION	• US : Large, sustained downturn in GDP starting from 1929 • UK : Avoided a banking crisis, experienced a modest fall in GDP
GOLDEN AGE	• US : Technology leader • Outside US : Diffusion of technology creates catch-up growth, improving productivity
FINANCIAL CRISIS	• US : Housing bubble creates banking crisis • Germany, Nordic countries, Japan, Canada, Australia : Did not experience bubble, largely avoided financial crisis
INTERNATIONAL OPENNESS (ALL THREE PERIODS)	More important in most countries than in the US

Figure 17.5b *The Great Depression, the golden age, and the financial crisis in cross-national comparison: Distinctive features of the United States.*

17.2 THE GREAT DEPRESSION, POSITIVE FEEDBACKS, AND AGGREGATE DEMAND

Capitalism is a dynamic economic system and, as we saw in Unit 12, booms and recessions are a recurrent feature even when weather-driven fluctuations in agricultural output are of limited importance in the economy. But not all recessions are equal. In Unit 13, we saw that in 1929 a downturn in the US business cycle similar to others in the preceding decade transformed into a large-scale economic disaster—the Great Depression.

THE GREAT DEPRESSION

The period during the 1930s in which there was a sharp fall in output and employment, experienced in many countries.

- Countries that left the *gold standard* earlier in the 1930s recovered earlier.
- In the US, Roosevelt's *New Deal* policies accelerated recovery from the Great Depression, partly by causing a change in expectations.

The story of how the *Great Depression* happened is dramatic to us, and must have been terrifying to those who experienced it. Small causes led to ever-larger effects in a downward spiral, like the cascading failures of an electricity grid during a blackout. Three simultaneous positive feedback mechanisms brought the American economy down in the 1930s:

- *Pessimism about the future:* The impact of a decline in investment on unemployment and of the stock market crash of 1929 on future prospects spread fear among households. They prepared for the worst by attempting to save more, bringing about a further decline in consumption demand.
- *Failure of the banking system:* The resulting decline in income meant that loans could not be repaid. By 1933, almost half of the banks in the US had failed, and access to credit shrank. The banks that did not fail raised interest rates as a hedge against risk, further discouraging firms from investing and curbing household spending on automobiles, refrigerators and other durable goods.
- *Deflation:* Prices fell as unsold goods piled up on store shelves.

Deflation affects aggregate demand through several routes. The most important channel operated through the effect of deflation on those with high debts. This positive feedback channel was new because in earlier episodes of deflation levels of debt had been much lower. Households stopped buying cars and houses, and many debtors become insolvent, creating problems for both borrowers and the banks. A survey showed that one-fifth of those in owner-occupied and rented accommodation was in default. Farmers were among those with high levels of debt: prices of their produce were falling, pulling down their incomes directly and pushing up the

burden of their debt. They responded to this by increasing production, which made the situation worse. When prices are falling, people also postpone the purchase of durables, which further reduces aggregate demand.

DISCUSS 17.1: FARMERS IN THE GREAT DEPRESSION

The response of farmers may have made sense from an individual point of view, but collectively it made the situation worse. Use diagrams, for example the model of a firm in a price-taking market for an individual farm and diagrams for supply and demand for the industry (for example wheat), to show why.

Few understood these positive feedback mechanisms at the time, and the government's initial attempts to reverse the downward spiral failed. This was partly because the government's actions were based on mistaken economic ideas. It was also because, even if they had pursued ideal policies, the government share of the economy was too small to counter the powerful destabilising trends in the private sector.

Figure 17.6 shows the fall in industrial production that started in 1929. In 1932 it was less than 60% of the 1929 level. This was followed by a recovery, until it fell again by 20% in 1937. Unemployment did not fall below 10% until 1941, the year the US entered the second world war. Consumer prices fell with GDP from 1929 to 1933 and remained stable until the early 1940s.

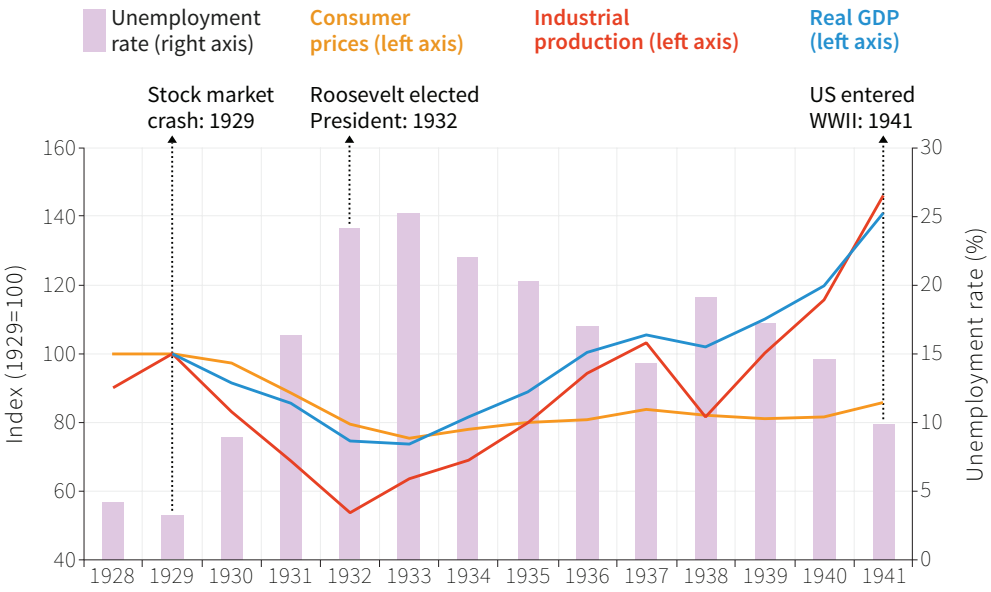


Figure 17.6 *The effect of the Great Depression on the US economy (1928-1941).*

Source: *United States Bureau of the Census. 2003. Historical Statistics of the United States: Colonial Times to 1970, Part 1. United States: United States Govt Printing Office; Federal Reserve Bank of St Louis (FRED).*

17.3 POLICYMAKERS IN THE GREAT DEPRESSION

Australia experienced a Black Saturday. The origin of the Great Depression can be dated to a day now known as Black Thursday. On Thursday 24 October 1929 the US Dow Jones Industrial Average plunged by 11% during the day, starting three years of decline for the US stock market. Figure 17.7 shows the business cycle upswings and downswings from 1924 to 1941.

Figure 17.7 *Changes in the components of aggregate demand during upswings and downswings (Q3 1924 to Q4 1941).*

Source: *Appendix B in Gordon, Robert J. 1986. The American Business Cycle: Continuity and Change. Chicago, Il: University of Chicago Press.*

The long downswing from the third quarter of 1929 until the first quarter of 1933 was driven by big falls in household and business investment (the red bar) and in consumption of non-durables (the green bar). Recall that in Figure 13.6 we used the multiplier model to describe how this shock created a fall in aggregate demand, and in Figure 13.8 we described a model of how households had cut consumption to restore their target wealth, to understand the observed behaviour of households and firms in the Great Depression.

In Unit 13, we showed how government policy could both amplify and dampen fluctuations. In the opening years of the Great Depression, government policy both amplified and prolonged the shock. Initially, government purchases and net exports hardly changed. As late as April 1932 President Herbert Hoover told Congress that "far-reaching reduction of governmental expenditures" were necessary, and advocated a balanced budget. Hoover was replaced by Franklin Delano Roosevelt in 1932, at which point government policy changed.

Fiscal policy in the Great Depression

Fiscal policy made little contribution to recovery until the early 1940s. Estimates suggest that output was 20% below the full employment level in 1931, for example, which means that the small budget surplus in that year would have implied a large cyclically adjusted surplus, given the decline in tax revenues in the depressed economy.

Under Roosevelt, from 1932 to 1936 the government ran deficits. When the economy went into recession in 1938-39, the deficit shrank from its peak of 5.3% in 1936 to 3% in 1938. This was another mistake that reinforced the downturn. The big increase in military spending from early 1940 (well before the US entered the second world war in late 1941) contributed to the recovery.

Monetary policy in the Great Depression

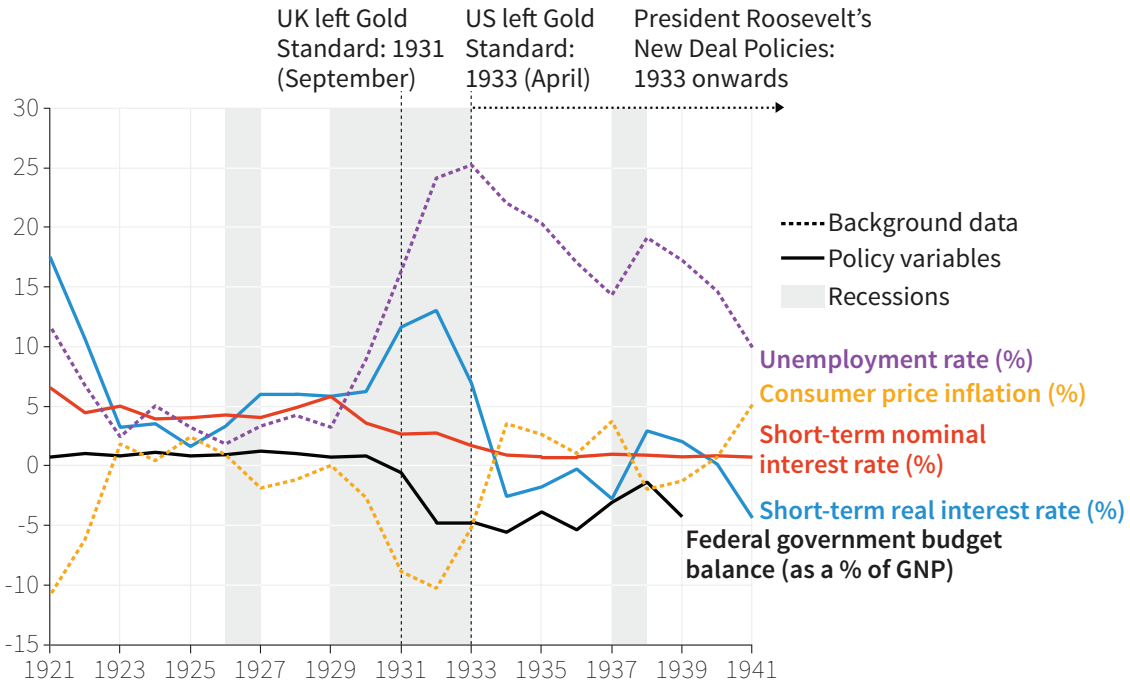


Figure 17.8 *Policy choices in the Great Depression: United States (1921-1941).*

Source: *Friedman, Milton, and Anna Jacobson J. Schwartz. 1982. Monetary Trends in the United States and the United Kingdom, Their Relation to Income, Prices, and Interest Rates, 1867-1975. Chicago, Il: University of Chicago Press; United States Bureau of the Census. 2003. Historical Statistics of the United States: Colonial Times to 1970, Part 1. United States: United States Govt Printing Office; Federal Reserve Bank of St Louis (FRED).*

Monetary policy prolonged the Great Depression. The real interest rate data in Figure 17.8 suggest that monetary policy was contractionary in the US economy from 1925 onwards: the real interest rate increased, reaching a peak of 13% in 1932. Once the downturn began in 1929, this policy stance reinforced, rather than offset, the decline of aggregate demand. But note that the nominal interest rate was falling after its peak in 1929; the real interest rate went up because prices were falling too. Interest-sensitive spending on buildings and consumer durables decreased sharply.

The gold standard

The US was still on what was known as the *gold standard*. This meant that the US authorities promised to exchange dollars for a specific quantity of gold (the promise was to pay an ounce of gold for $20.67). Under the gold standard, the authorities had to continue to pay out at the fixed rate and, if there was a fall in demand for US dollars, gold would flow out of the country. To prevent this, either the country's tradable goods must become more competitive (boosting gold inflows through higher net exports) or gold must be attracted through capital inflows by putting up the nominal interest rate, or keeping it high relative to the interest rate in other countries. As a result, policymakers were reluctant to push the interest rate down to the *zero lower bound* to avoid contributing to the gold outflow.

Unless wages decline rapidly to raise international competitiveness and boost the inflow of gold through higher exports and lower imports, sticking to the gold standard in a recession is destabilising: it will amplify the downturn. There was a very large outflow of gold from the US after the UK left the gold standard in September 1931. One reason for speculation against the US dollar—that is, investors selling dollars for gold—was that there were expectations that the US would also abandon the gold standard and devalue the dollar. If it did, those holding dollars would lose.

Countries that left the gold standard earlier in the 1930s recovered earlier.

A change in expectations

In 1933 Roosevelt began a programme of changes to economic policy:

- The *New Deal* committed federal government spending to a range of programmes to increase aggregate demand.
- The US left the gold standard in April 1933, which meant the US dollar was devalued to $35 per ounce of gold, and the nominal interest rate was reduced to close to the zero lower bound (see Figure 17.8).
- Roosevelt also introduced reforms to the banking system following the bank runs of 1932 and early 1933.

The change in people's *beliefs* about the future was just as important as these policy changes. On 4 March 1933, in his inaugural address as president, Roosevelt had told Americans that: "the only thing we have to fear is fear itself—nameless, unreasoning, unjustified terror".

We have seen that the terrors of consumers and investors in 1929 had been justified. But a combination of Roosevelt's New Deal policies and the beginnings of recovery in the economy that were already underway before he became president, households and firms began to think that prices would stop falling and that employment would expand.

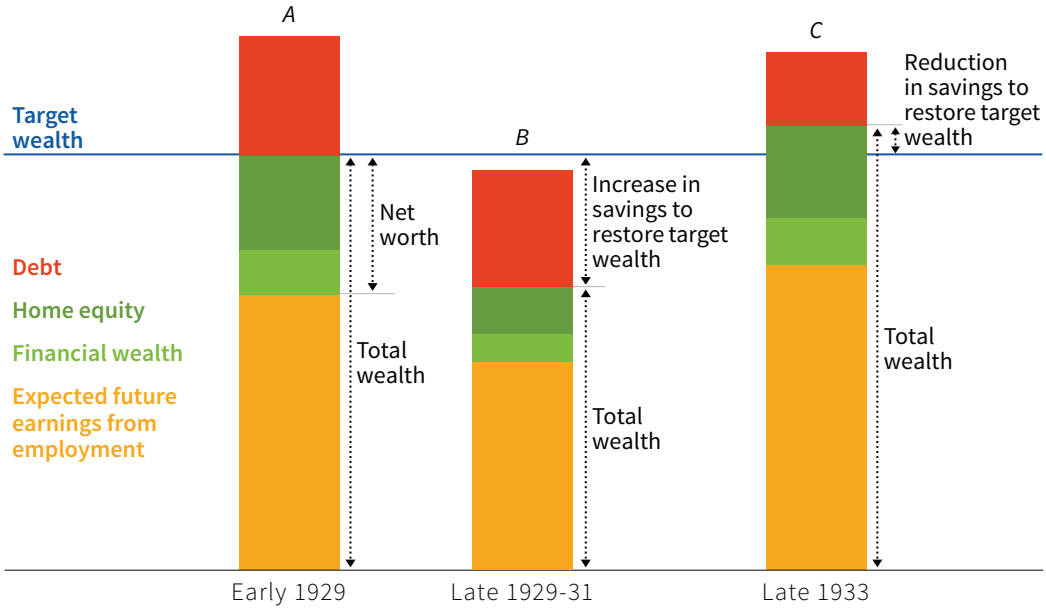


Figure 17.9 *The Great Depression: Households cut consumption to restore target wealth.*

Figure 17.9 adds a third column to the model that we first encountered in Figure 13.8. Column C shows the household's perspective from late 1933. By that time output and employment were growing. With much of the uncertainty about the future resolved, households re-evaluated their expected wealth (including their expected earnings from employment). They reversed the cutbacks in consumption because they saw no need to make additional savings. To the extent that they now expected their income prospects and asset prices to return to pre-crisis levels, consumption would be restored. Any increase in wealth above target due to the increased savings during the Depression years (shown by wealth above target in column C) would create an additional boost to consumption.

The slow path to recovery had begun. But the US economy would not return to pre-Depression levels of employment until Roosevelt was in his third term as president and the second world war had begun.

17.4 THE GOLDEN AGE OF HIGH GROWTH AND LOW UNEMPLOYMENT

The years from 1948 until 1973 were remarkable in the history of capitalism. In the US, we saw in Figure 17.2 that productivity growth was more rapid and unemployment was lower than in the other periods. But this 25-year *golden age* was not confined to the US. Countries across western Europe, Japan, Australia, Canada and New Zealand experienced a golden age as well. Unemployment rates were historically low (see Figure 15.1). Figure 17.10 shows data from 1820 to 1913 for 13 advanced countries, and for 16 countries from 1950.

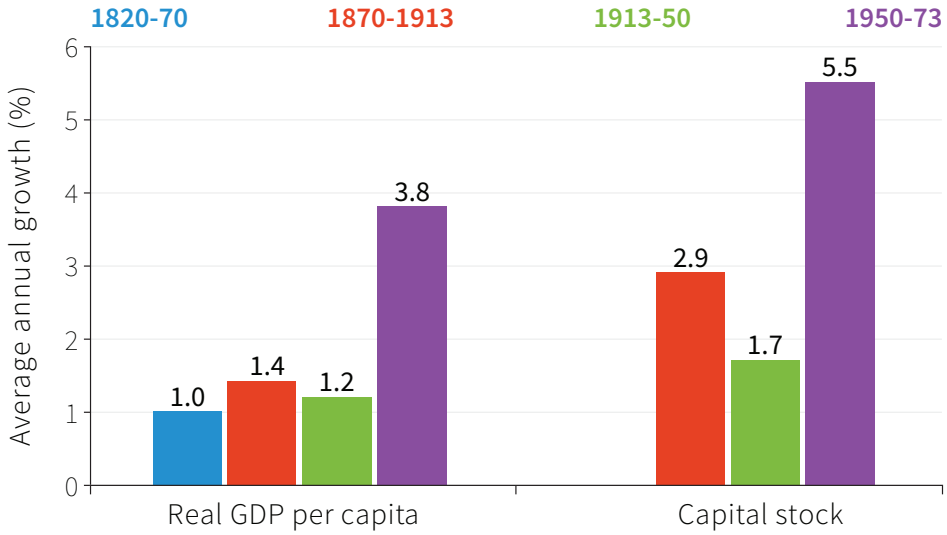


Figure 17.10 *The golden age of capitalism in historical perspective.*

Source: *Table 2.1 in Glyn, Andrew, Alan Hughes, Alain Lipietz, and Ajit Singh. 1989. 'The Rise and Fall of the Golden Age.' In The Golden Age of Capitalism: Reinterpreting the Postwar Experience, edited by Stephen A. Marglin and Juliet Schor. New York, NY: Oxford University Press.*

The growth rate of GDP per capita was more than two-and-a-half times as high during the golden age than in any other period. Instead of doubling every 50 years, living standards were doubling every 20 years. The importance of saving and investment is highlighted in the right panel, where we can see that the capital stock grew almost twice as fast during the golden age as it did between 1870 and 1913.

The story of how the large western European countries and Japan (almost) caught up to the US is told in Figure 17.11. In the figure, the level of GDP per hour worked in the US is set at the level of 100 throughout, and so the figure tells us nothing about the performance of the US itself (we have to use Figure 17.2 for that). However, it is a striking way to represent the starting point of these economies relative to the US immediately after the second world war and their trajectories in the years that followed. This was known as *catch-up growth*.

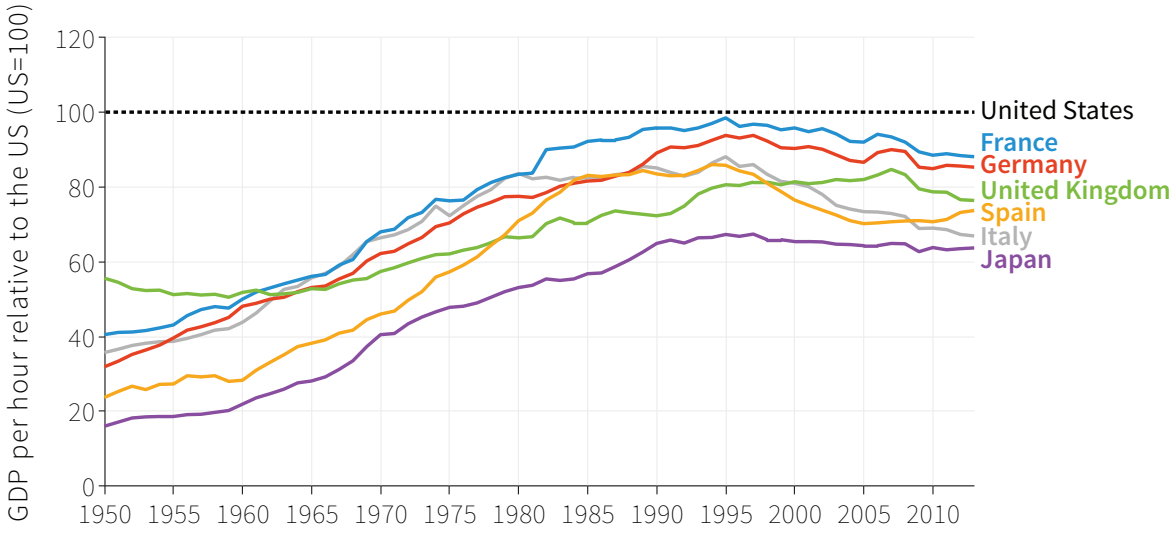


Figure 17.11 *Catch-up to the US during the golden age and beyond (1950-2013).*

Source: *The Conference Board. 2014. 'Total Economy Database.'*

The three large defeated countries (Germany, Italy and Japan) were furthest behind in 1950. Japan's GDP per hour worked was less than one-fifth the level of the US. Clearly, growth of all of these economies was faster than the US during the golden age: all moved much closer to the level of US productivity.

What was the secret of golden age performance in the productivity leader—the US—and in the follower countries?

- *Changes in economic policymaking and regulation:* These resolved the problems of instability that characterised the Great Depression
- *New institutional arrangements between employers and workers:* These created conditions in which it was profitable for firms to innovate. In the US, the technology leader, this meant new technologies, while the follower countries often adopted improved technology and management already in use in the US. Because workers' trade unions and political parties were now in a stronger position to bargain for a share of the productivity gains, most supported innovation—even when it meant temporary job destruction.

THE GOLDEN AGE OF CAPITALISM

The period of high productivity growth, high employment and stable inflation extending from the end of the second world war to the early 1970s.

- The gold standard was replaced by the more flexible *Bretton Woods System.*
- Employers and employees shared the benefits of technological progress thanks to the *postwar accord*.
- The golden age ended with a period of *stagflation* in the 1970s.

After the second world war governments had learned the lessons of the Great Depression. This affected national and international policymaking. Just as Roosevelt's New Deal signalled a new policy regime and raised expectations in the private sector, postwar governments provided reassurance that policy would be used to support aggregate demand if necessary.

Government was now larger in all of these countries after the second world war, and the size of government grew throughout the 1950s and 1960s. Figure 13.1 showed the decline in output fluctuations after 1950, and the much larger size of government in the US. In Unit 13, we saw how a larger government provides more automatic stabilisation for the economy. The modern welfare state was built in the 1950s, and unemployment benefits were introduced. This also formed part of the automatic stabilisation.

Given the cost of adherence to the gold standard during the Great Depression, it was clear that a new policy regime for international economic relations had to be put in place. The new regime was called the *Bretton Woods System* after the ski resort in New Hampshire where representatives of the major economies, including Keynes, created a system of rules that was more flexible than the gold standard. Exchange rates were tied to the US dollar rather than gold and, if countries became very uncompetitive—if they faced a "fundamental disequilibrium" in external accounts, in the words of the agreement—devaluations of the exchange rate were permitted. When a currency like the British pound was devalued (as occurred in November 1967) it became cheaper to buy pounds. This boosted the demand for British exports and reduced the demand of British residents for goods produced abroad. The Bretton Woods System worked fairly well for most of the golden age.

17.5 WORKERS AND EMPLOYERS IN THE GOLDEN AGE

High investment, rapid productivity growth, rising wages and low unemployment defined the golden age.

This seems too good to be true. We saw a model of the wage and profit curve in Unit 15 which highlighted the conflict of interest between workers and employers: at low unemployment, workers must get high wages so that they will work effectively. This depresses profits and reduces investment. The golden age does not seem to follow this model: we saw low unemployment, high profits and high investment at the same time.

How did this virtuous circle work?

- *Profits after taxes in the US economy remained high:* This persisted from the end of the second world war through the 1960s (look again at Figure 17.3) and the situation was similar in other advanced economies.
- *Profits led to investment:* The widespread expectation that high profits would continue in the future provided the conditions for sustained high levels of investment (refer back to the model of investment spending in section 13.4).
- *High investment and continued technological progress created more jobs:* Unemployment stayed low.
- *The power of workers:* Trade unions and political movements allied with employees had high bargaining power, which allowed a sustained increase in wages.

As the last bullet suggests, trade unions were important in this process, as well as governments. Between 1920 and 1933 trade unions lost two-fifths of their members, most of the losses occurring immediately after the first world war. During the 1930s changes in the laws affecting trade unions, as well as the hardship of the Great Depression, reversed this decline. High demand for labour during the second world war strengthened labour's bargaining power: trade union membership as a fraction of total employment peaked in the early 1950s. There was a subsequent steady decline during the next 50 years.

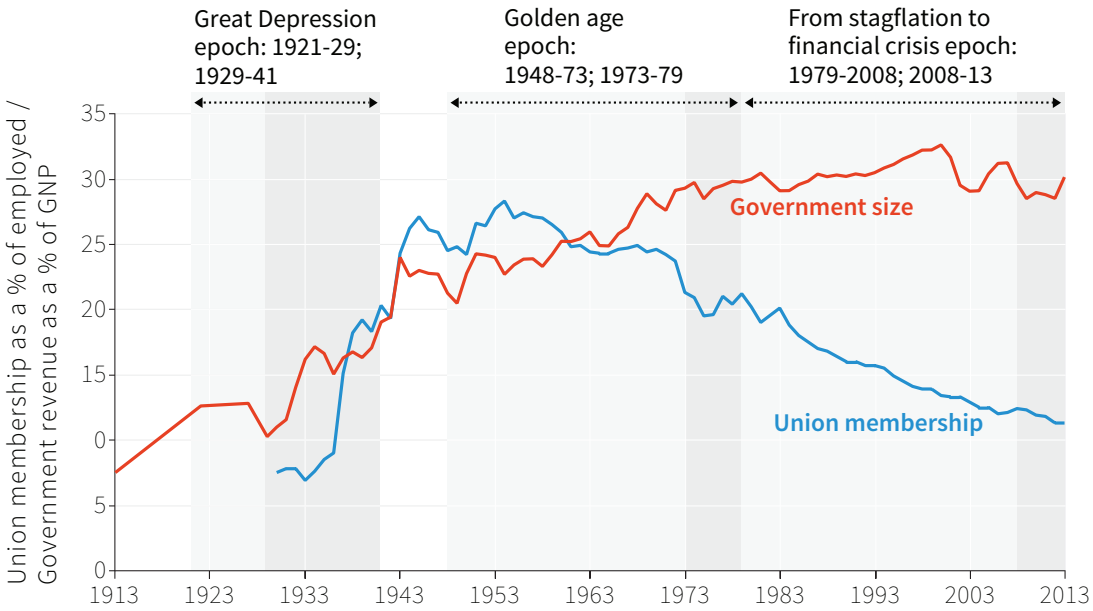


Figure 17.12 *Trade union membership and the size of government in the United States (1913-2013).*

Source: *Wallis, John Joseph. 2000. 'American Government Finance in the Long Run: 1790 to 1990.' Journal of Economic Perspectives 14 (1): 61–82; Mayer, Gerald. 2004. Union Membership Trends in the United States. Washington, DC: Congressional Research Service; US Bureau of Economic Analysis.*

Figure 17.12 shows both the growth of the government and the historically high level of trade union membership in the US. As we have seen, larger government partly reflected the new unemployment insurance entitlement. From the wage and profit

curve model, we know that higher unemployment benefits and stronger trade unions shift the wage curve upwards, allowing employees to bargain for a share of increasing productivity.

In the golden age employees had sufficient bargaining power to claim a share in the gains that technological progress made possible. Both employers and employees realised that there was more to be gained in cooperating to increase the size of the pie than in wasting resources in futile efforts to claim most of the pie for themselves. Policies, business practices and trade union strategies during this period reflected this insight.

When translated into the labour market model (in Figure 17.13) the four bullets explaining of the golden age can be translated into shifts in the profit curve and the wage curve:

- *The profit curve shifted up:* This happened because productivity increased rapidly
- *The wage curve shifted up:* Low unemployment, strong unions and favourable government policies increased labour's bargaining power, but the resulting upward shift in the wage curve was modest, allowing for high profits, high investment (the basis of continuing productivity growth) and low unemployment.

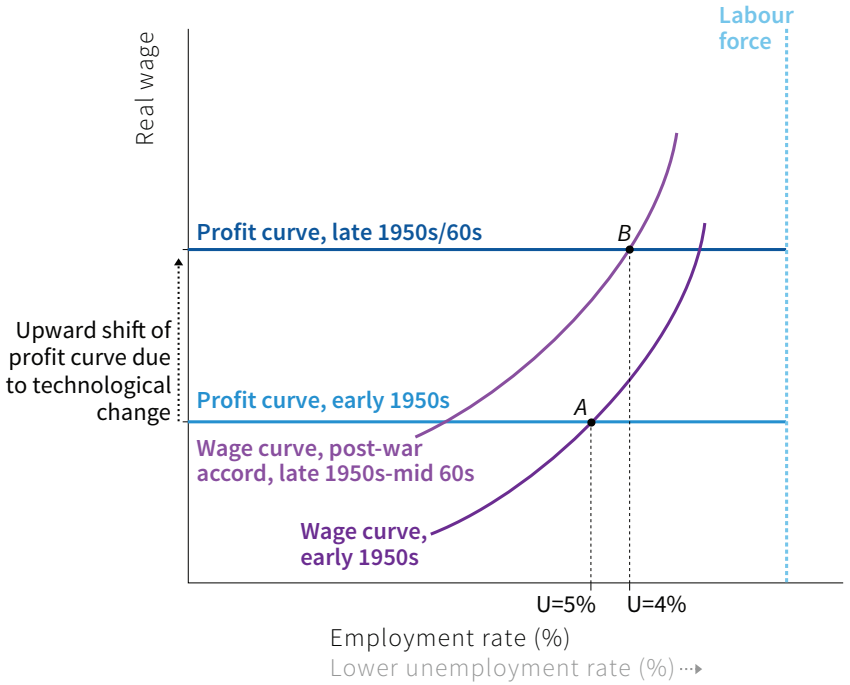


Figure 17.13 *The golden age: Using the wage and profit curves.*

Recall from Unit 15 that the profit curve shows the real wage consistent with employers maintaining investment at a level to keep employment constant. This means that a real wage above the profit curve will drive firms to leave (relocate to some other economy) or cut back on their investment, and employment falls.

The profit curve will shift up when worker productivity rises, or when taxation on profits is reduced, or when investors and owners have optimistic expectations about future profits. It will shift down when employers have to pay higher prices for imported raw materials, such as oil.

In the US, technological progress was rapid in the golden age as the innovations developed during the Great Depression and the second world war were embodied in new capital equipment.

The new technologies and new management techniques already in use in the US could also be used in the catch-up economies if the innovators expected high enough profits. In many of these countries golden age growth was even faster than at the technology frontier as defined by the US in Figure 17.11.

Taking the example of the US, we can represent the economy as at point A at the beginning of the golden age, with unemployment of 5%. Technological progress shifts the profit curve up (to the one labelled "late 1950s/60s").

Unless wages adjust upwards or the economy expands, the result initially is a wage much below the profit curve. This stimulates high investment, consistent with the data for the growth of the capital stock in the US shown in Figure 17.3.

But wages eventually did rise and at the same time the economy expanded, moving towards point *B* in the figure.

The strength of unions in wage setting and the improvement in unemployment insurance during the 1950s and 60s are illustrated as an upward shift of the wage curve in Figure 17.13. To get the outcome observed, with wages growing in line with productivity at low unemployment such as point *B*, unions and employers need to agree about the scope for wage increases. This would be the case if the wage curve shifted to the one labelled "Wage curve, postwar accord, late 1950s-mid 60s".

Unions would refrain from using the full extent of their bargaining power (for example, in firms or plants where they had a very strong position) and cooperate in an economy-wide bargain designed to keep wage growth consistent with the constraint imposed by the profit curve. In return, employers would maintain investment at a level sufficient to keep unemployment low. This unwritten but widely observed pattern of sharing the gains to technological progress between employees and employers is termed the *postwar accord*. In Unit 15, we also referred to this process as *fair-shares bargaining*.

Different countries had different postwar accord relationships among employers, unions and governments to create high productivity growth, high real wage growth and low unemployment. In Scandinavia, Austria, Belgium, Netherlands, Switzerland and West Germany, wage setting was either centralised in a single union, or coordinated among unions or employers' associations, resulting in wage restraint. In

technologically advanced sectors in France and Italy, governments intervened to set wages in dominant state-owned firms, creating wage guidance across the economy. The outcome was similar to the result in the countries with centralised wage setting.

Where there was little cooperation between employers and unions, a country's performance in the golden age was worse. In Figure 17.11, the UK's relatively poor golden age performance shows up clearly: it started with higher productivity than the other large countries shown (that is, its productivity level in 1950 was the closest to that of the US) but was overtaken by France, Italy and Germany in the 1960s.

Compared to Sweden, Norway and many continental European nations, where the postwar accord underwrote rapid growth in productivity, the British industrial relations system made an accord difficult. It combined very strong union power at the factory level with fragmented unions, which were unable to cooperate in the economy as a whole. The strength of local union shop stewards (representatives), in a system of multiple unions per plant, led unions to attempt to outdo each other when negotiating wage deals, and created opposition to the introduction of new technology and new ways of organising work. The problems of British firms were compounded because markets in former British colonies were protected from competition.

Competition is important in the Schumpeterian creative destruction process because it creates incentives for firms to get a step ahead of the competition, and reduces the number of low-productivity firms. When competition is weak, existing firms and jobs are protected. The employers and workers in these firms share the monopoly rents, but the overall size of the pie is reduced because technological progress is slower.

Postwar accords succeeded in the US and the successful catch-up countries in creating the conditions for a high profit and high investment equilibrium. It delivered rapid productivity and real wage growth at low unemployment, but the British experience during the 1950s and 1960s (Figure 17.11) emphasises that there is nothing automatic about achieving this outcome.

17.6 THE END OF THE GOLDEN AGE

The virtuous circle of the golden age began to break down in the late 1960s, as a result of its own successes. The postwar accord and its rationale of enlarging the pie gave way to a return to contest over the size of the slice that each group could get. This set the stage for the period of combined inflation and stagnation called stagflation that would follow. Employers eventually won the contest, but at a substantial cost to the economy.

Australia can go many years without a major bush fire. But fewer small fires means there will be more flammable undergrowth, which increases the chance of a major fire. Years of low unemployment (fewer small fires) convinced workers that they had little fear of losing their jobs. Their demands for improvements in working conditions and higher wages drove down the profit rate.

They also demanded policies to redistribute income to the less well off and to provide more adequate social services, making it difficult for governments to run a budget surplus. In the US, additional military spending to fund the Vietnam war added to aggregate demand, keeping the economy at unsustainably high levels of employment.

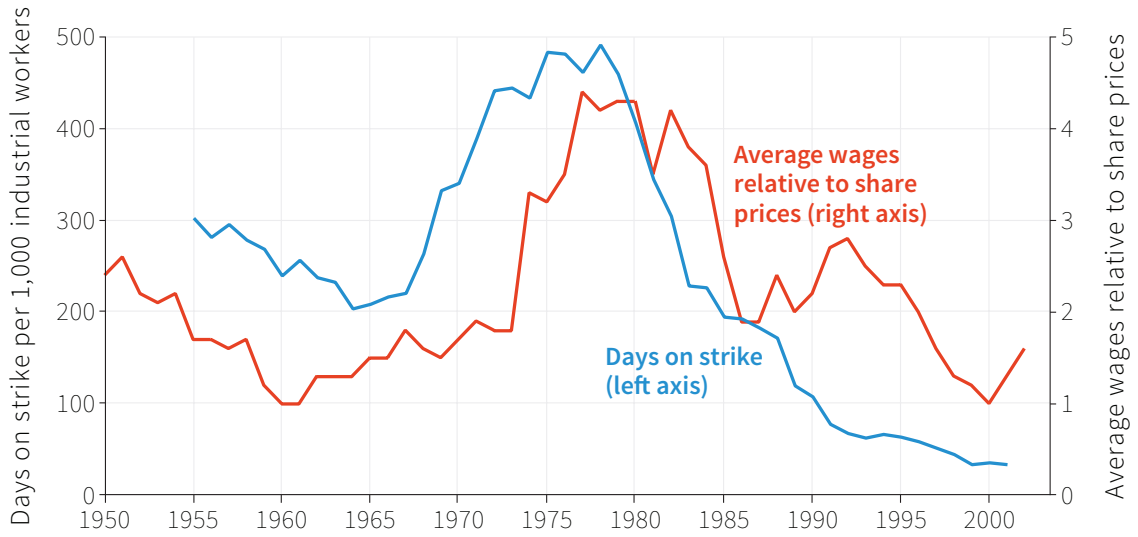
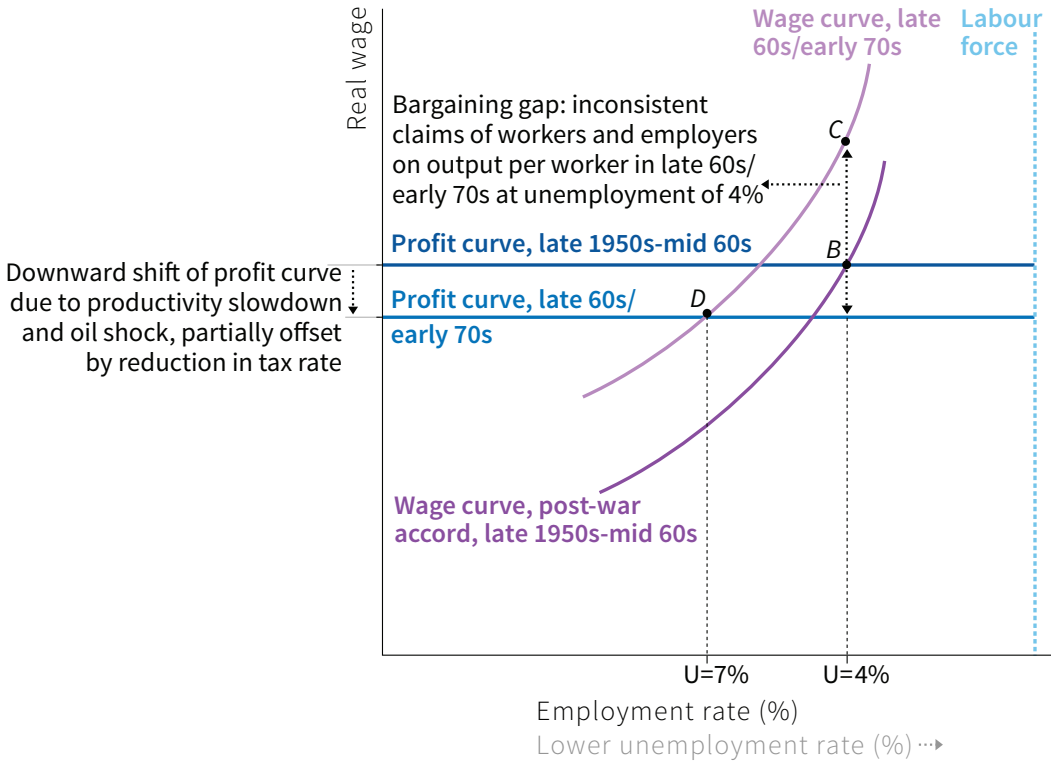


Figure 17.14 *The end of the golden age: Strikes and wages relative to share prices in advanced economies (1950-2002).*

Source: *Glyn, Andrew. 2006. Capitalism Unleashed: Finance, Globalization, and Welfare. Oxford: Oxford University Press.*

Greater industrial strife in the late 1960s signalled the breakdown of the golden age accords. Figure 17.14 plots the days on strike per 1,000 industrial workers in advanced economies from 1950 to 2002. As strike activity peaked, wages measured relative to share prices increased rapidly. The postwar accords that helped create the golden age collapsed.

The process is represented in Figure 17.15 by an upward shift in the wage curve (to the one labelled "late 60s/early 70s"). At the same time, economy-wide productivity growth slowed (see Figure 17.2 for the US data). In the catch-up countries in western Europe, it was becoming more difficult to get easy gains from technology transfer, because the gap between US technology and the technology used by followers narrowed (see Figure 17.11). In 1973, the first oil price shock occurred. In the Figure 17.15, this pushes the profit curve down (see the profit curve labelled "1973-79").



The combination of a downward shift in the profit curve and an upward shift in the wage curve meant that the sustainable long-term unemployment rate increased to 7%, shown at point *D*. The double-headed arrow at low unemployment shows the situation in the early 1970s.

Figure 17.15 *The end of the golden age: Using the wage and profit curves.*

In the early 1970s the claims of employers given their bargaining power compared to consumers (the profit curve) and the claims of workers given their bargaining power compared to their employers were no longer consistent. Something had to give. The golden age was over.

What happened?

Wages did not rise to the level of point C. Under the impact of the upward pressure on wages and the oil price shock, the economy contracted and unemployment began to rise. But even a significant reduction in the employment rate (short of increasing the unemployment rate to 7%) did not eliminate the bargaining gap shown in the figure. A result was an increase in the rate of inflation, as is shown in Figure 17.16.

Because of the strong bargaining position of workers in the early 1970s in most of the high-income economies, the oil price shock primarily hit employers, redistributing income from profits to wages (Figure 17.15). The era of fair-shares bargaining was coming to a close.

In the US, where trade unions were less powerful, workers nevertheless managed to defend their share of the pie even after the oil price increase. In countries with inclusive and powerful trade unions (as described in Unit 15), the accord survived. In Sweden, for example, the powerful centralised labour movement restrained its wage claims to preserve profitability, investment and high levels of employment.

But in virtually all countries including the US, wages remained above the new profit curve, so investment fell and the rate of productivity growth slowed. As predicted by the model in Figure 17.15, the outcome was rising inflation (Figure 17.16) falling profits (Figure 17.3), weak investment (Figure 17.3), and high unemployment (Figure 17.16).

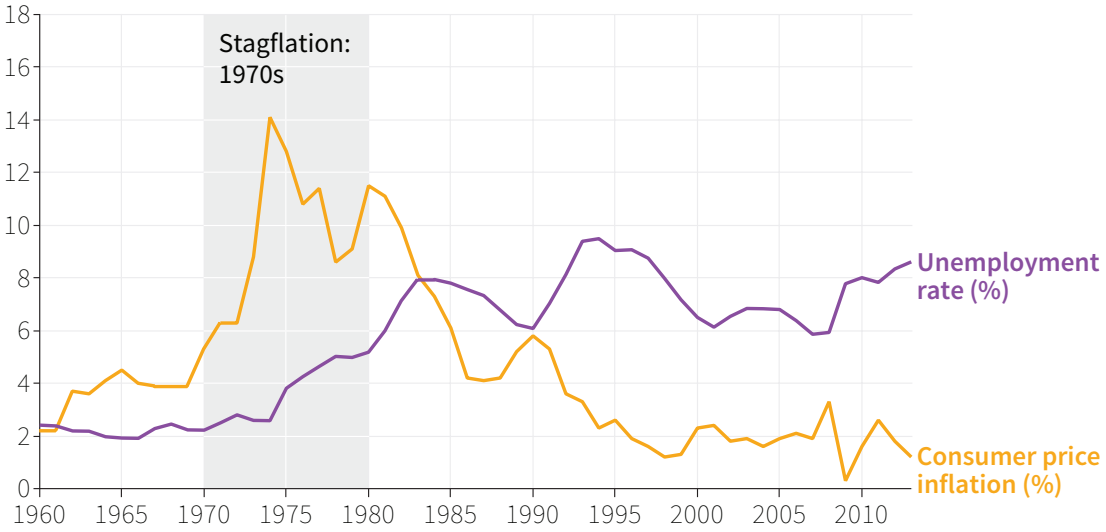


Figure 17.16 *After the golden age: Unemployment and inflation in advanced economies (1960-2013).*

Source: *OECD. 2015. 'OECD Statistics.'*

The end of the golden age set off a new economic crisis—one that was very different from the Great Depression. The economic downturn of the 1930s had been propelled by problems of aggregate demand and for this reason it has been called a *demand-side* crisis. The end of the golden age has been called a *supply-side* crisis, because those problems on the supply side of the economy depressed the profit rate, the rate of investment and the rate of productivity growth.

The period that ensued came to be called stagflation because it combined high unemployment and high inflation. If the golden age was an unusual time during which everything went right at once, stagflation was the unusual time when everything went wrong.

According to the Phillips curve model of Unit 14, inflation goes up when unemployment goes down; this is a movement along the Phillips curve. Figure 17.16 summarises the unemployment and inflation data for the advanced economies from 1960-2013.

Just as the Phillips curve predicts, for most of the period, inflation and unemployment were negatively correlated: as unemployment rose, inflation fell and vice versa. But the entire Phillips curve shifted upward during this period period, as a bargaining gap opened and expected inflation increased. Look at the shaded part of Figure 17.16: inflation and unemployment rose together, giving this period its name.

17.7 AFTER STAGFLATION: THE FRUITS OF A NEW POLICY REGIME

The third major epoch during the last 100 years of capitalism began in 1979. Across the advanced economies, policymakers focused on restoring the conditions for investment and job creation. Expanding aggregate demand would not help: what would have been part of the solution during the Great Depression had now become part of the problem.

Arrangements based on accords between workers and employers continued in northern European and Scandinavian countries. Elsewhere, employers abandoned the accord, and policymakers turned to different institutional arrangements as the basis for restoring the incentives for firms to invest.

The new policies were called *supply-side reforms*, aimed to address the causes of the supply-side crisis of the 1970s. The policies were centred on the need to shift the balance of power between employer and worker in the labour market, and in the firm. Government policy at this time achieved this goal in two main ways:

- *Restrictive monetary and fiscal policy:* Governments showed that they were prepared to allow unemployment to rise to unprecedented levels, weakening the position of workers and restoring the consistency of claims on output as the basis of modest and stable inflation.
- *Shifting the wage curve down:* As we saw in Unit 15, these policies included cuts in unemployment benefits and the introduction of legislation to reduce trade union power.

Figure 17.16 illustrates the new policy environment. Unemployment increased rapidly from 5% to 8% in the early 1980s. This was the price of restoring conditions for profit and investment, and for reducing inflation from greater than 10% to 4%. Policymakers were prepared to depress aggregate demand and tolerate high unemployment until inflation fell.

DISCUSS 17.2: WORKERS' BARGAINING POWER

After the Great Depression most advanced economies adopted policies that strengthened the bargaining power of employees and labour unions. After the golden age, by contrast, the policies weakened workers' bargaining power.

1. Explain the reasons for these contrasting approaches.
2. With hindsight, do you think the economic logic behind each set of policies makes sense?

The increased unemployment beginning with the first oil price shock in 1973 had two effects:

- It reduced the bargaining gap in Figure 17.15, bringing down inflation (shown in Figure 17.16).
- It put labour unions and workers on the defensive as the cost of job loss rose and employees' bargaining power eroded.

Figure 17.17 shows the development of productivity (output per hour) and real wages in manufacturing in the US from the beginning of the golden age. Index numbers are used for each series to highlight the growth of real wages relative to that of output per hour worked. Real wage growth in line with output per hour is not inevitable: in Unit 1, when looking at the growth of real wages in England since the 13th century, we saw that institutions (social movements, changes in the voting franchise and in laws) played a vital role in translating productivity growth into real wage growth.

The figure shows two dramatically different periods:

- *Before 1973*: Fair-shares bargaining meant that wages and productivity grew together.
- *After 1973*: Productivity growth was not shared with workers. For production workers in manufacturing, real wages barely changed in the 40 years after 1973.

By the mid-1990s, the effects of the new supply-side policy regime were becoming clear. The period from this time until the global financial crisis of 2008 was called the great moderation because inflation was low and stable, and unemployment was falling. Although wage growth fell well below productivity growth, policymakers no longer thought of this as a bug; it was a feature of the new regime. The third oil shock that occurred in the 2000s was a good test of the regime. As we saw in Unit 14, it created none of disruption of the two oil shocks in the 1970s.

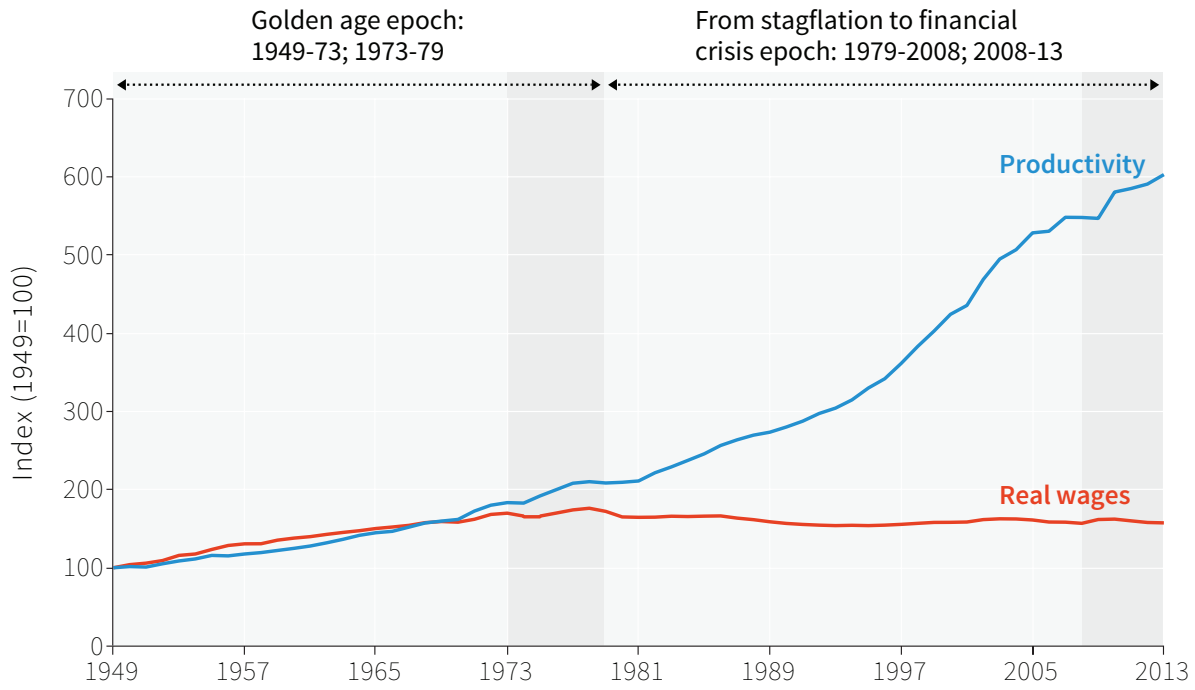


Figure 17.17 *The Golden Age and its Aftermath: Real wages and output per production worker in manufacturing in the United States (1949-2013).*

Source: *US Bureau of Labor Statistics. Note: "production workers" exclude supervisory employees such as foremen and managers.*

In virtually all of the advanced economies the new supply-side policies redistributed income from wages to profits. In the US (Figure 17.3) the after-tax profit rate gradually increased between the 1970s and 2008. But investment responded only weakly to the profit incentives, so that the rate of growth of the capital stock declined. The economy had settled at some point below the profit curve in Figure 17.15, with more than sufficient profits to motivate an expansion to the higher employment equilibrium, but with investment not fully responding.

Supply-side policy advisors could not recreate the improbable package of high employment, high investment and growing wages of the golden age. The growth of profits unmatched by investment in new equipment would also help to cause the next crisis.

17.8 BEFORE THE FINANCIAL CRISIS: HOUSEHOLDS, BANKS AND THE CREDIT BOOM

The great moderation masked three changes that would create the environment for the global financial crisis. While these changes were common across the advanced economies, actors in the US economy played a pivotal role in the global financial crisis, just as they had during the Great Depression:

- *Rising debt:* The sum of the debt of the government and of non-financial firms changed relatively little as a proportion of GDP between 1995 and 2008, but the mountainous shape of total debt in the US economy shown in Figure 17.4 was created by growth in household and financial sector debt.
- *Increasing house prices:* Rising house prices, which became more pronounced after 1995.
- *Rising inequality:* The long-run decline in inequality that began after the Great Depression reversed after 1979 (Figure 17.2). Workers no longer shared in the gains from productivity.

How can we make an argument that connects the financial crisis to the great moderation, and to long-run rising debt, house prices and inequality? We use what we learned in Units 9, 11, 12 and from section 17.4 to help us. We know that, during the great moderation, from the mid-1990s to the eve of the financial crisis, the real wages of those with earnings in the bottom 50% hardly grew. Relative to the earnings of the top 50%, they lost out. One way they could improve their consumption possibilities was to take out a home loan. Before the 1980s, financial institutions had been restricted in the kinds of loans they could make and in the interest rates they could charge. *Financial deregulation* generated aggressive competition for customers, and gave those customers much easier access to credit.

THE GREAT MODERATION AND THE GLOBAL FINANCIAL CRISIS

The *great moderation* was a period of low volatility in output between the 1980s and 2008. It was ended by the *global financial crisis*, triggered by falling US house prices from 2007 onwards.

- At the onset of the crisis, government and central bank stabilisation policies, notably including *bank bailouts*, avoided a repeat of the Great Depression.
- Nevertheless, there followed a sustained global fall in aggregate output, popularly known as the *great recession*.

Housing booms and the financial accelerator

When households borrow to buy a house, this is a secured or collateralised loan. As part of the mortgage agreement, the bank can take possession of the house if the borrower does not keep up repayments. Collateral plays an important role in sustaining a house price boom. When the house price goes up—driven, for example, by beliefs that a further price rise will occur—this increases the value of the household's collateral (see the left-hand diagram in Figure 17.18). Using this higher collateral, households can increase their borrowing, and move up the housing ladder to a better property. This, in turn, pushes up house prices further; because the banks extend more credit based on the higher collateral, it sustains the bubble. Increased borrowing, made possible by the rise in the value of the collateral, is spent on goods and services as well as on housing.

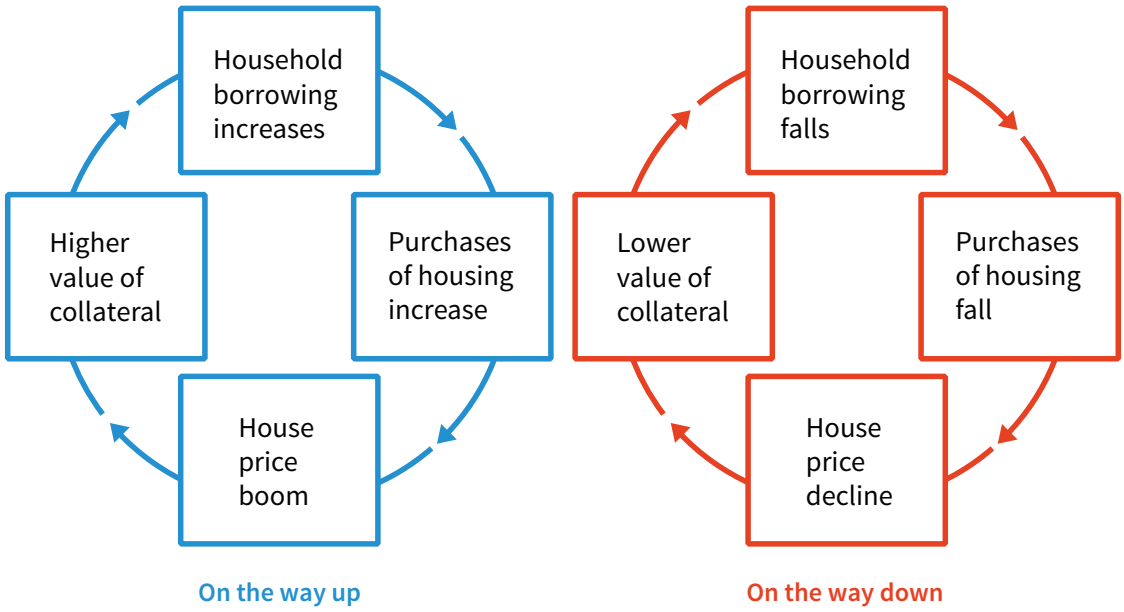


Figure 17.18 *The housing market on the way up and on the way down.*

Source: *Adapted from figure in Shin, Hyun Song. 2009. 'Discussion of "The Leverage Cycle" by John Geanakoplos'. Also presented as Figure 6.1 in Carlin, Wendy, and David Soskice. 2014. Macroeconomics: Institutions, Instability, and the Financial System. Oxford: Oxford University Press.*

When house prices are expected to rise, it is attractive to households to increase their borrowing. Suppose a house costs $200,000, and the household makes a downpayment of 10% ($20,000). This means it borrows $180,000. Its initial *leverage ratio*, in this case the value of its assets divided by its equity stake in the house, is $200/20 = 10$. Suppose the house price rises 10% to $220,000. The return to the equity the household has invested in the house is 100% (since the value of the equity stake has risen from $20,000 to $40,000: it has doubled). Households that are convinced that house prices will rise further will want to increase their leverage: that is how they get a high return. The increase in collateral, due to the rise in the price of their house, means they can satisfy their desire to borrow more.

The mechanism through which a rise in the value of collateral leads to an increase in borrowing and spending by households and firms is called the financial accelerator (look this up in Unit 13 if you cannot remember the details). The left-hand side of Figure 17.18 shows the outcome of the interaction between the bubble in house prices and its transmission through the economy via the financial accelerator during a boom. On the right-hand side, we see what happens when house prices decline: the value of collateral falls and the household's spending declines, pushing house prices down.

The assets and liabilities of a household can be represented in its balance sheet, and this can be used to explain the interaction of a house price bubble and the financial accelerator. The house is on the asset side of the household's assets. The mortgage owed to the bank is on the liabilities side. When the market value of the house falls below what is owed on the mortgage, the household has negative net worth. This condition is sometimes referred to as the household being "underwater". In the example: if the leverage ratio is 10, a fall in the house price by 10% wipes out the household's equity.

As we saw with households in the Great Depression, if a decline in net worth means that a household is below its target wealth, it responds by cutting what it spends. When a housing bubble is forming, the rise in the value of collateral reinforces the boom by boosting both borrowing and spending; on the way down, the fall in the value of house increases household debt and the household reduces spending. Rising house prices immediately before 2008 were prices that sent the "wrong" message. We know that resources were misallocated because the US, and some countries in Europe, were left with thousands of abandoned houses.

DISCUSS 17.3: LAGGING BEHIND THE WEALTHY

1. Use section 10.9 to identify the market failure described in the quote below.
2. Would you recommend policy intervention to correct this market failure?

"In 1995 [Mr Baggett] moved into a house in the Harvard-Yale section of Salt Lake, [US], a tree-lined neighbourhood near the University of Utah that is home to many doctors, lawyers and professors. Mr Baggett used credit cards to furnish the home with the kind of carpets and furniture his neighbours and relatives could afford. 'I felt insecure; I was an hourly-paid worker in this fancy neighbourhood,' says Mr Baggett. He says he was making $13 an hour for a time doing back-office work at a local bank while supporting two children."
– Wall Street Journal, *Lagging Behind the Wealthy, Many Use Debt to Catch Up* (2005)

Financial deregulation and subprime borrowers

In the boom period the expectation of rising house prices reduced the riskiness of home loans to the banks making them and, as a result, banks extended more loans. The opportunities for poor people to borrow for a home loan expanded as lenders asked for lower deposits, or even no deposit at all. This is shown in Figure 17.19. The financial accelerator mechanism is an example of positive feedback: from higher collateral, to more borrowing, to further increases in house prices.

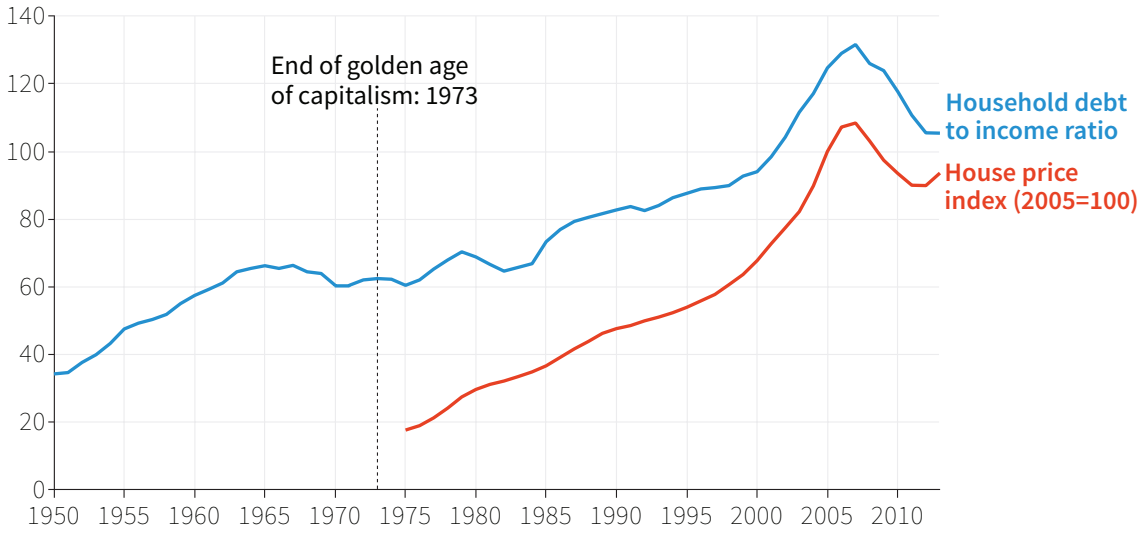


Figure 17.19 *The household debt-to-income ratio and house prices in the United States (1950-2014).*

Source: *US Federal Reserve. 2015. 'Financial Accounts of the United States, Historical.' December 10; US Bureau of Economic Analysis; Federal Reserve Bank of St Louis (FRED).*

Figure 17.20 shows the contrast between the material wealth of a household in the bottom and top fifth of households, according to their net worth in 2007. Using the definitions introduced in section 13.3 and used in section 17.4, the household's material wealth is equal to the value of its house (which will by definition be equal to the sum of the debt outstanding and the household's home equity) minus the mortgage debt, plus financial wealth (net of non-housing debt).

The left-hand bar represents borrower households. These are poor households, normally only able to borrow when they have housing collateral to use as security. They have little financial wealth, as shown by the size of the green rectangle. These households have much more debt than equity in their houses, and are vulnerable to a fall in house prices.

Rich households have a lot of assets, mainly in the form of financial wealth: bank account and money market deposits, government and corporate bonds, and shares. They also have little debt. These are the saver households of Unit 11.

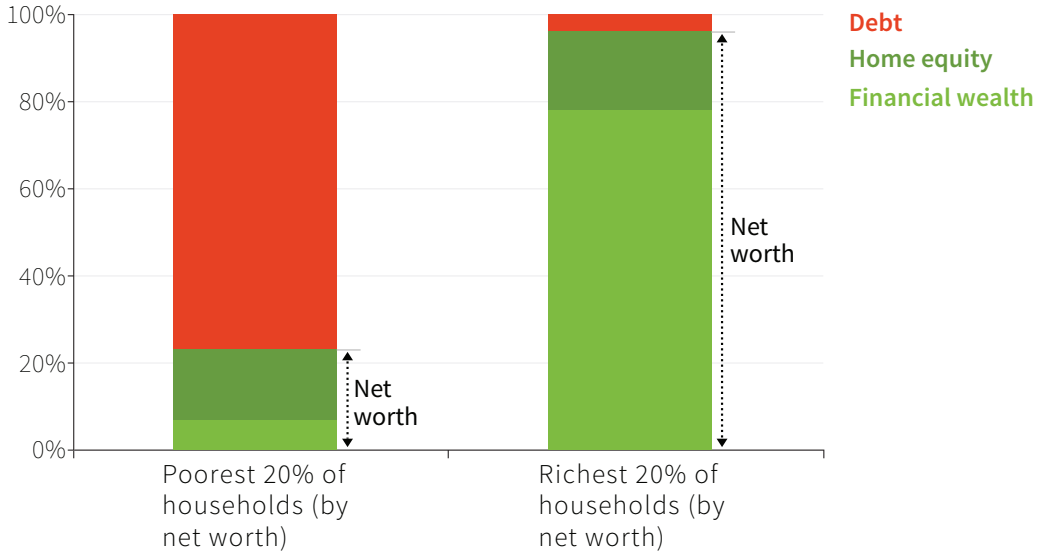


Figure 17.20 *Household wealth and debt in the United States: Poorest and richest quintiles by net worth, 2007.*

Source: *Adapted from Figure 2.1 in Mian, Atif, and Amir Sufi. 2014. House of Debt: How They (and You) Caused the Great Recession, and How We Can Prevent It from Happening Again. Chicago, Il: The University of Chicago Press.*

DISCUSS 17.4: HOUSEHOLD WEALTH AS A BALANCE SHEET

1. Show the information in Figure 17.20 in the form of example balance sheets for one household from the lowest and one from the highest net worth quintile (use the balance sheet of the bank in section 11.10 as a guide).

Think about the proportions of debt held by these households that might consist of mortgage debt. Now consider the relative effects on the households of a fall in house prices.

2. In your example balance sheet for the poorer family, estimate the fall in house prices that would push this household into negative equity.
3. Would such a household be insolvent? Explain.

Financial deregulation and bank leverage

In the context of the deregulated financial system, banks increased their borrowing:

- To extend more loans for housing
- To extend more loans for consumer durables like cars and furnishings
- *To buy more financial assets based on bundles of home loans*

The combination of the great moderation, rising house prices, and the development of new, apparently less risky, financial assets such as the *derivatives* called *collateralised debt obligations* (CDOs), based on bundles of home loans called *mortgage-backed securities* (MBSs), made it profitable for banks to become more highly leveraged.

Figure 17.21 shows the leverage of US investment banks, and of all UK banks:

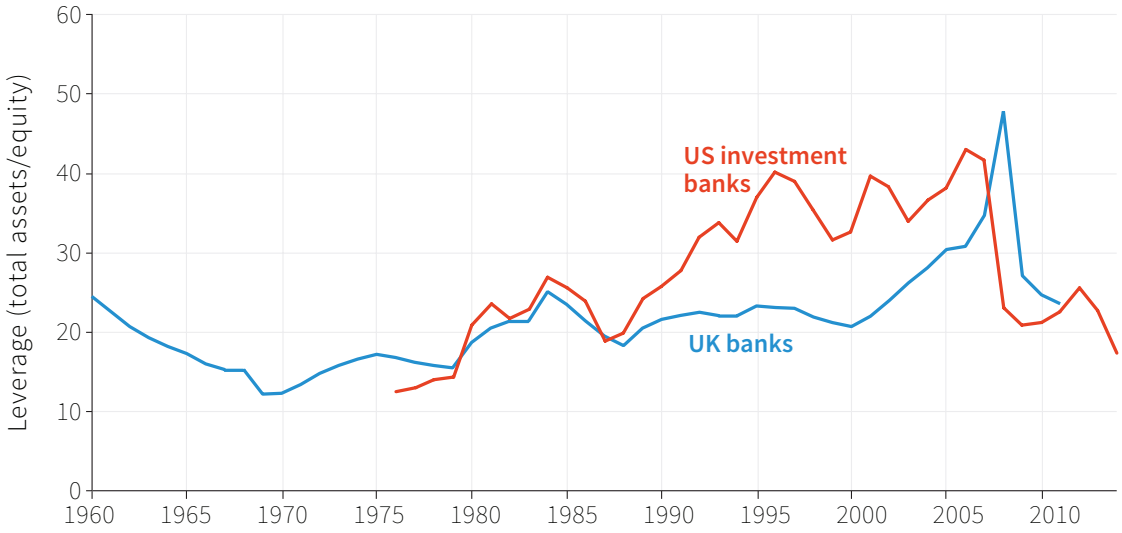


Figure 17.21 *Leverage of banks in the UK and US (1960-2014).*

Source: *US Federal Reserve. 2015. 'Financial Accounts of the United States, Historical.' December 10; Bank of England. 2012. Financial Stability Report, Issue 31.*

In the US, the leverage of investment banks was between 12 and 14 in the late 1970s, rising to more than 30 in the early 1990s. It hit 40 in 1996 and peaked at 43 just before the financial crisis. By contrast, the leverage of the median UK bank remained at the level of around 20 until 2000. Leverage then increased very rapidly to a peak of 48 in 2007. In the 2000s British and European global banks, including firms called shadow banks, increased borrowing to buy CDOs and other financial assets that originated in the US housing market.

Leverage increased because of financial deregulation and the business model of banks. But why were savers prepared to continue lending to the increasingly leveraged financial system and, indirectly, to the highly leveraged household sector?

Firms called *credit ratings agencies* (the big three are Fitch Ratings, Moody's and Standard & Poor's) assess the risk of financial products, and part of their role is to provide evidence to reassure lenders that their investments are safe. After almost 20 years of the great moderation economic crises seemed like a historical idea, and so these companies gave the highest ratings (meaning the lowest risk) to many of the assets created from *subprime mortgages*.

The subprime housing crisis of 2007

The interrelated growth of the indebtedness of poor households in the US and global banks meant that, when homeowners began to default on their repayments in 2006, the effects could not be contained within the local or even the national economy. The crisis caused by the problems of subprime mortgage borrowers in the US spread to other countries. Financial markets were frightened on 9 August 2007 when French bank BNP Paribas halted withdrawals from three investment funds because it could not "fairly" value financial products based on US mortgage-based securities: it simply did not know how much they were worth.

The recession that swept across the world in 2008-09 was the worst contraction of the global economy since the Great Depression. Unlike the bushfires in south-eastern Australia in 2009, the financial crisis took the world by surprise. The world's economic policymakers were unprepared. They discovered that a long period of calm in financial markets could make a crisis more likely.

This was an argument that the economist Hyman Minsky had made long before the great moderation. Minsky developed these ideas while a professor of economics at the University of California, Berkeley, and so he may even have been thinking of fires: in northern Mexico the fire management authorities allow small fires to burn, and as a result dry undergrowth does not accumulate. Major fires are more frequent across the US border in California, where small fires are quickly extinguished.

In 1982 Minsky wrote a book about the way in which tranquil conditions lead firms to choose riskier methods of financing their investment. His warning went unheeded. Instead of producing increased vigilance, the calm conditions of the great moderation bred complacency among regulators and economists. It was the increasingly risky behaviour of banks, as Minsky had predicted, that created the conditions for the crisis.

GREAT ECONOMISTS

HYMAN MINSKY

Hyman Minsky (1919-1996) was an American economist who developed a financial theory of the business cycle. His ideas have attracted renewed attention among both academics and banking and finance professionals since the global economic crisis of 2008.

Minsky argued that macroeconomic fluctuations could not be properly understood without taking account of the manner in which business investment is financed. At a time when most economists viewed firms as the location of a production function, Minsky focused instead on the assets and liabilities on the firm's balance sheet. The assets, including plant and equipment, but also less tangible assets such as patents, copyrights and trademarks, are expected to generate a steam of revenues stretching far into the future. The liabilities include the firm's obligations to its creditors, and imply a stream of payments due at various points in time.

New investment by the firm expands its capacity to produce goods and services, and thus alters its expected stream of revenues. If it is financed by debt, it also changes the firm's financial obligations at future dates. In deciding how to finance its investment, the firm faces a choice:

- *Issue long-term debt:* It anticipates that revenues would be sufficient to cover obligations at all points in time.
- *Issue short-term debt:* This debt needs to be repaid before the anticipated revenues are available. It creates the need for further borrowing to repay debt at the end of this term.

In general long-term borrowing is more expensive, since lenders demand a higher interest rate. But short-term borrowing is risky, because the firm may be unable to refinance debt as it comes due. Even if it can refinance, it may be forced to borrow at high rates if credit availability is constrained.

Firms that chose the safer but more expensive option, matching revenues and debt obligations, were said by Minsky to be engaged in *hedge finance*. Those that took the cheaper but more risky option, borrowing short-term to finance long-term investments, were engaging in *speculative finance*.

A key component of Minsky's theory concerned the manner in which the distribution of financial practices in the economy changed over time. As long as financial market conditions remained relatively tranquil, so that rolling over short-term debt was easy, firms with the most aggressive financial practices would prosper at the expense of those that were the most prudent. Not only would the most aggressive firms grow faster, they would also attract imitators, and the distribution of financial practices in the economy would become increasingly speculative. There would be a rise in the demand for refinancing short-term debt, and hence an increase in financial fragility: a severe financial market disruption, with a contraction in credit or a spike in short-term interest rates, would become increasingly likely.

In Minsky's view, this process leads inevitably to a crisis because, as long as a crisis is averted, the most aggressive financial practices proliferate and financial fragility continues to rise. When a crisis finally occurs, the most aggressive firms will suffer disproportionately and the prudent firms will prosper. The sharp shift in the aggregate distribution of financial practices lowers fragility and sets the stage for the process to begin again. In Minsky's words:

"Stability—even of an expansion—is destabilising in that more adventurous financing of investment pays off to the leaders, and others follow."
– Hyman Minsky, *John Maynard Keynes* (1975)

A period like the great moderation, in other words, sows the seeds of the next financial crisis.

Some echoed Minsky's thinking. In September 2000, Sir Andrew Crockett, general manager at the Bank for International Settlements, told banking supervisors:

"The received wisdom is that risk increases in recessions and falls in booms. In contrast, it may be more helpful to think of risk as increasing during upswings, as financial imbalances build up, and materialising in recessions."
— Andrew Crockett, *Marrying the Micro- and Macro-Prudential Dimensions of Financial Stability* (2000).

In 2007 Charles Prince, chief executive of Citigroup, explained to the *Financial Times* the difficulty of resisting "adventurous financing" during booms. "As long as the music is playing, you've got to get up and dance," he said in July, as the global economy hurtled towards a crisis deeper than anything seen since the Great Depression, "We're still dancing."

17.9 THE FINANCIAL CRISIS AND THE GREAT RECESSION

Rising house prices in the US in the 2000s were driven by the behaviour of lenders, encouraged by government policy, to extend loans to poorer households. They were able to fund these subprime loans by packaging them into financial derivatives, which banks and financial institutions across the world were eager to buy. Rising house prices created the belief that prices would continue to rise, which shifted the demand curve further to the right by providing households with access to loans based on housing collateral.

Figure 17.22 illustrates the house price cycle on the way down. The housing market in the US economy in 2006 is shown at point A. Once prices began to fall, the demand curve for housing shifted to the left. It was apparent to households that housing was no longer an asset that could be counted on to increase in value. This was the shift from A to B: it can be related to the fall in the house price index, from a level of 100 at the peak to 92 in 2007. Once prices were falling, the belief took hold that prices would fall further. The demand curve shifted further to the left: the house price index fell to 76 in 2008.

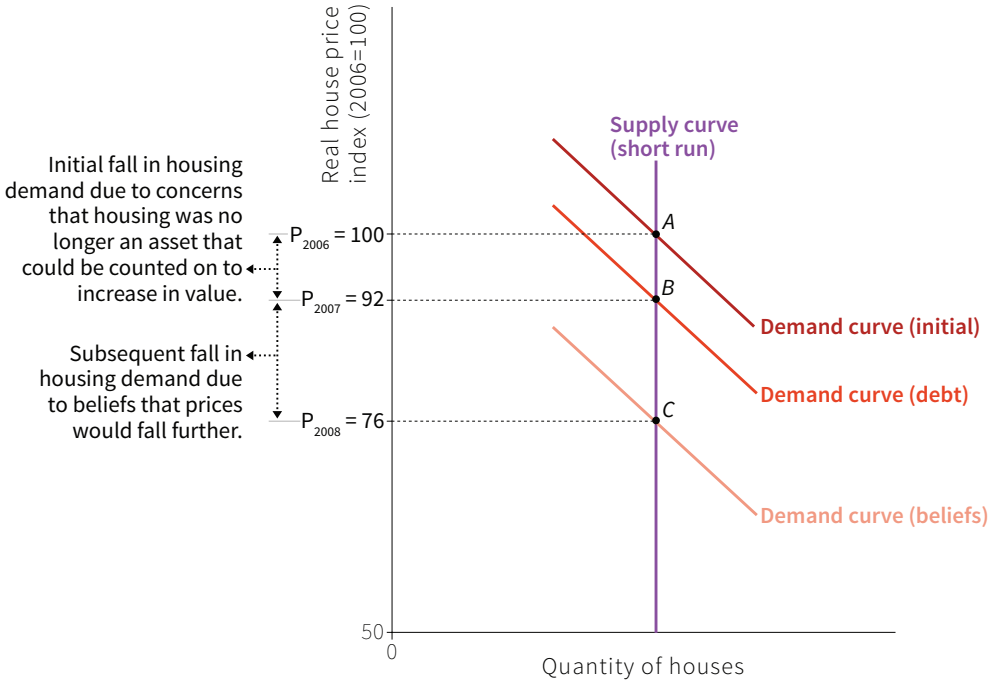


Figure 17.22 *The financial crisis: House price collapse in the US.*

Source: *Bank for International Settlements. 2015. 'Residential Property Price Statistics.' November 20, and other national sources.*

In Figure 17.23 you can see the contribution of the components of GDP to growth in the 18 months before the crisis in the US economy, then in the five quarters of recession from the start of 2008, followed by the recovery phase to the end of 2010. The fall in residential investment (the solid red bar) was the most important feature of the onset phase: at that stage it was the only drag on growth. This was the consequence of the fall in house prices that began in 2006. In the recession, a further fall in housing investment was compounded by a fall in non-residential investment and consumption.

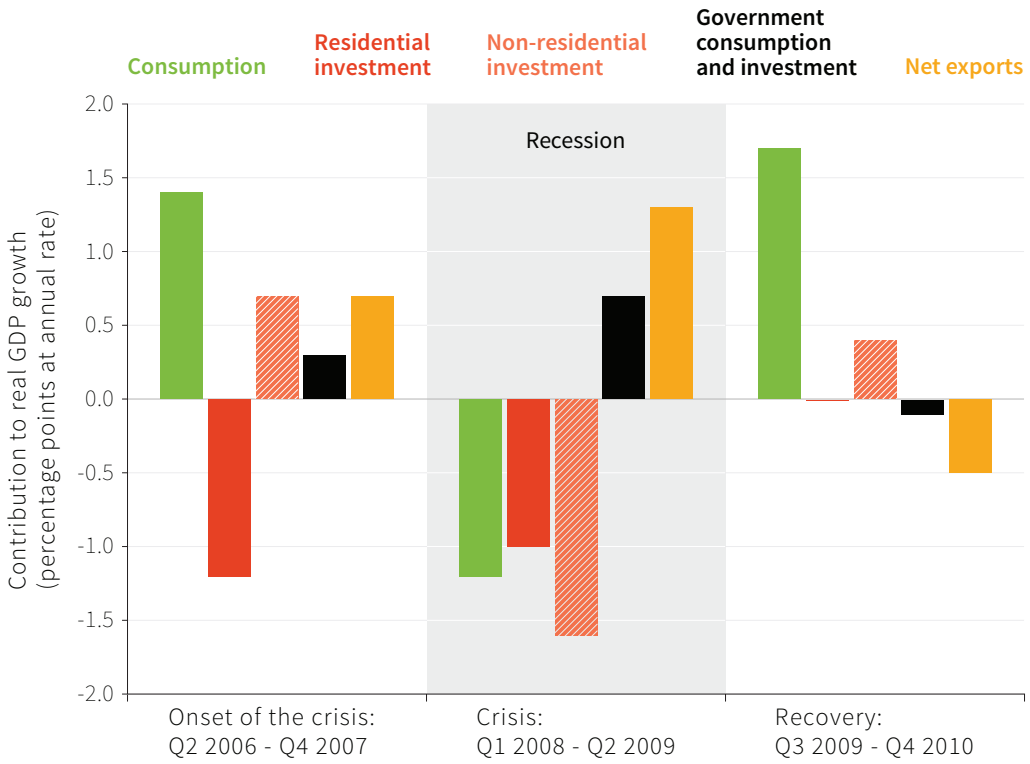
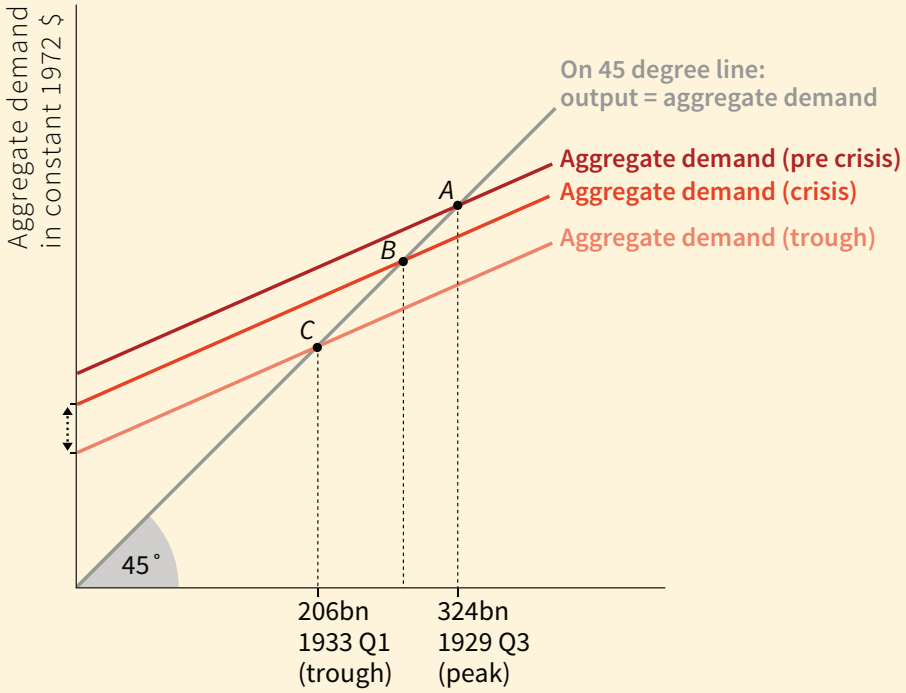


Figure 17.23 *Aggregate demand and the financial crisis in the US (Q2 2006 to Q4 2010).*

Source: *US Bureau of Economic Analysis.*

Just as in the Great Depression, the fall in consumption was not simply due to the multiplier process. Households stopped buying new houses, but also cut spending on consumer durables. The financial accelerator mechanism helps to explain the transmission of falling house prices through the fall in the value of collateral to aggregate demand. Cutbacks in spending on new housing and on consumer durables were concentrated among the poorer households who had taken out subprime mortgages. The timing of the collapse of demand is consistent with the central role played by housing and debt in the financial crisis. There was also a fall in investment. Orders for new equipment were cancelled and factories closed. Workers were laid off; job creation slumped.

DISCUSS 17.5: THE CRISIS AND THE MULTIPLIER



The US economy in 1929
The aggregate demand line defines a level of goods market equilibrium before the crisis. In Q3 of 1929, output was $324bn, its highest level.

Decline in aggregate demand (late 1929 to early 1930)
The fall in firm and household investment created an initial fall in aggregate demand.

The US economy in early 1930
The new goods market equilibrium was at point *B*.

The US economy in 1933
The downward shift in the consumption and investment functions in 1930 and 1931, associated with uncertainty about earnings and assets, the banking crisis, falling prices and higher real interest rates meant that aggregate demand continued to fall. By 1933 output had declined from $324bn to $206bn.

Aggregate demand in the Great Depression: Multiplier and positive feedback processes.

1. Show the features of the 2008 crisis in the multiplier diagram using the figure above for the Great Depression as a model. Use the concepts of the consumption function, a house price bubble, the financial accelerator and positive feedback in your answer.
2. How can you represent the role played by the higher marginal propensity to consume of households in the bottom quintile in your analysis?

We can link the pattern of aggregate demand in Figure 17.23 to the decisions of households by using a diagram similar to the one we developed for the Great Depression. This is Figure 17.24. These two figures are different ways of looking at the same developments: Figure 17.24 is an individual household-eye-view of the unfolding crisis; Figure 17.23 is the same process from the perspective of the whole economy.

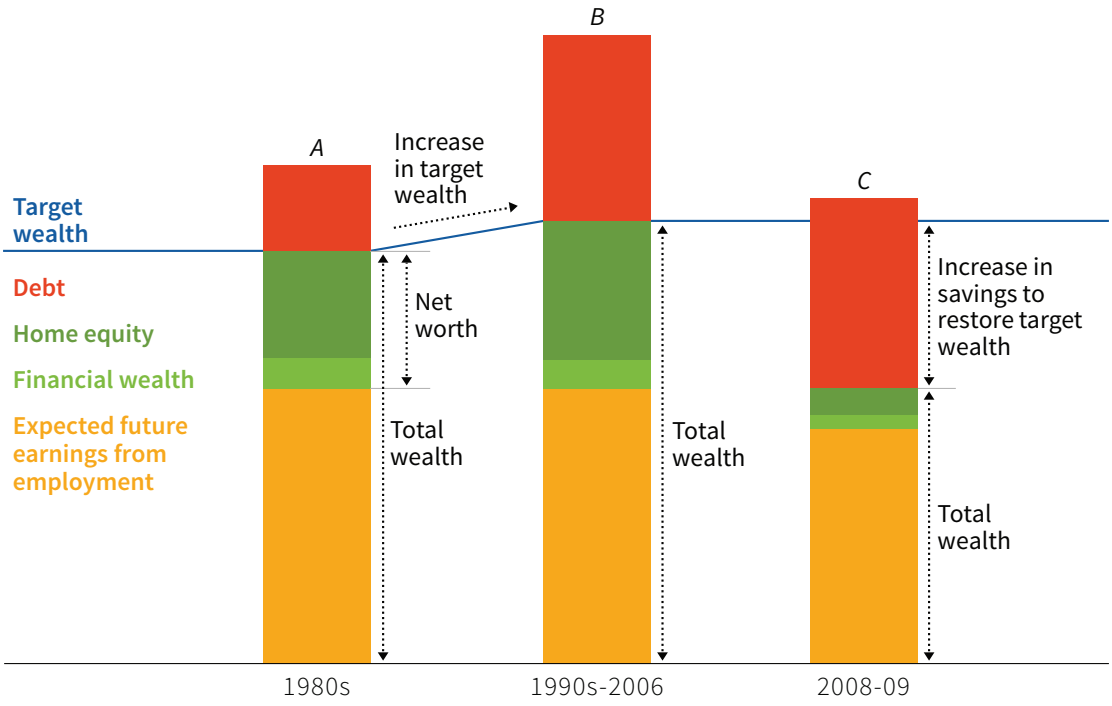


Figure 17.24 *The financial crisis: Housing boom, household debt and house price crash.*

Column A in Figure 17.24 shows the situation in the 1980s. As we have seen, through the 1990s and especially in the early 2000s, households took advantage of rising house prices to increase their debt. From Unit 11, we know that an increase in the price of a house allows the household to borrow more because the value of the collateral has risen. Over this period, expected earnings from employment remained constant for many households and we show this component of expected wealth as unchanged between column *A* and column *B* (the yellow rectangles). But the

increase in housing wealth pushed up households' assessment of their wealth (the dark green rectangles). The upward trend of house prices in real terms appeared to be a reliable feature of the world and households uprated their target wealth to reflect the increase in collateral and in their ability to borrow to raise their level of consumption. This is shown in column *B*. The higher level of debt of the household is shown by the larger debt rectangle (red).

From 2006, house prices in the US began to fall. The household's viewpoint in 2008 and 2009 is shown in column C. Rising unemployment led to a downward re-evaluation of expected future earnings from employment. Household net worth shrank, as we can see in column *C*. Note that the size of the debt rectangle has not changed between columns *B* and *C*. Given the fall in house and asset prices, the increased debt acquired in the boom years, combined with lower expected earnings, had the effect of reducing wealth below target. As a result, households cut consumption and increased savings. This is shown in column *C* by the double-headed arrow labelled "Increase in savings to restore wealth to target".

The household shown in Figure 17.24 still has positive net worth following the fall in house and asset prices in the crisis. This is shown by the sum of the red and green rectangles in column C. But the behaviour of households whose net worth became negative following the fall in house prices was an important feature of the great recession that followed the financial crisis in the US. To show this in a diagram like Figure 17.24, the debt rectangle would slide down into the box labelled "expected earnings from employment", wiping out the red and green rectangles, reducing total wealth, and increasing the gap between expected and target wealth. It is easy to see how households in the bottom quintile shown in Figure 17.20 went underwater in 2008 and 2009. In the US in 2011, 23% of properties with a mortgage were worth less than the mortgage. Households in this position would have cut consumption as they paid down their debt to restore their financial position.

17.10 THE ROLE OF BANKS IN THE CRISIS

House prices and bank solvency

The financial crisis was a banking crisis—and it was global, as BNP Paribas demonstrated in August 2007 when it would not pay out to bondholders in one of its investment funds. The banks were in trouble because they had become highly leveraged and were vulnerable to a fall in the value of the financial assets that they had accumulated on their balance sheets (refer back to Figure 17.21 for the leverage of US and UK banks). The values of the financial assets were in turn based on house prices.

With a ratio of net worth to assets of 4%, as in the example of the bank in Figure 11.15, a fall in the value of its assets of an amount greater than this will render a bank insolvent. House prices fell much more than 4% in many countries in the recent financial crisis. In fact, the peak-to-trough fall in house price indices for Ireland, Spain and the US were 50.3%, 31.6% and 34.6% respectively. This creates a problem of solvency for the banks: just as with the underwater households, banks were in danger of their net worth being wiped out. It is relatively easy for a household to calculate whether this has happened, but not for a bank.

Unlike a house, obscure financial assets on (or often designed to be kept off) a bank's balance sheet, with acronyms like *CDO, CDS, CLO* and even CDO^2, were hard to value. This made it difficult to judge which banks were in trouble.

Bank liquidity and the credit crunch

Doubts about the solvency of banks created another problem in the financial system—the problem of liquidity, which we introduced in Unit 11. A characteristic feature of banking is the mismatch between short-term liabilities, which it owes to depositors, and long-term assets, which are loans owed to the bank. In consequence banks rely on the money market to fund themselves when they need short-term liquidity. But the operation of the money market relies on borrowers and lenders having trust in the solvency of those with whom they trade. The expected profit on a loan is the interest rate multiplied by the probability that the borrower will not default:

$$\textit{Expected profitability of loan} = (1 + r) \times (1 - \textit{probability of default})$$

Therefore, as people feared that those to whom they were lending were more likely to default, they would only lend at a high interest rate. In many cases, banks or others operating in the money markets simply refused to lend at all. Newspapers called it the *credit crunch*.

In Unit 11 we learnt that the interest rate in the money market is tied tightly to the policy interest rate set by the central bank; this relationship broke down in the credit crunch. Borrowing on the interbank market became much more expensive and hampered the ability of central banks to stabilise the economy: even when the central banks reduced the interest rate to the zero lower bound, the fear that banks would default kept money market rates high. This led to high mortgage lending rates: high money market interest rates raised a bank's funding costs, as we would predict using the model in Unit 11.

Fire sales: A positive feedback process

The forced sale of assets, known informally as a *fire sale*, is a positive feedback process. It reinforces the fall in asset prices and hastens the insolvency of banks. In the financial crisis, the fire-sale external effect affected both the housing market and the markets for financial assets, and both affected banks.

In the housing case, it is easy to visualise: think of a household that is underwater and cannot repay a housing loan. Its debt exceeds the market value of the house. The household defaults on repayments, and either walks out or is foreclosed by the bank. After foreclosure, the bank owns the house, which it sells. The bank accepts a low price because it is not in the business of owning and maintaining houses, and the value of the house falls further if it is unoccupied. This is a market failure due to the external effect of the fire sale, which is conferring a cost (a fall in price) on other owners of the same type of asset. Similar fire sales of financial assets by distressed banks push prices down, and impose costs on other asset owners by reducing their net worth. This in turn threatens their solvency.

Governments rescue banks

Across the advanced economies, banks failed and were rescued by governments (to find out how they did this, and for more background on how the financial system failed during the crisis, use this web site). In Unit 11 we highlighted the fact that banks do not bear all the costs of bankruptcy. The bank owners know that others (taxpayers or other banks) will bear some of the costs of the banks' risk-taking activity. So the banks take more risks than they would take if they bore all the costs of their actions. Excess risk-taking by banks is a negative external effect leading to a market failure. And it arises because of the principal-agent problem between the government (the principal) and the agent (the bank). The difference of interest arises because the government will bear the cost of bank bailout as a consequence of excessive risk-taking by banks. Governments cannot write a complete set of rules that would align the interests of the banks with those of the government or the taxpayer.

Banks are rescued because the failure of a bank is different from the failure of a typical firm or household in a capitalist economy. Banks play a central role in the payments system of the economy and in providing loans to households and to firms. Chains of assets and liabilities link banks, and those chains had extended across the world in the years before the crisis.

The interconnectedness of banks was vividly illustrated in the credit crunch, where liquidity dried up in the money markets because of doubts of each bank about the solvency of other banks. The event associated most closely with the financial crisis, the bankruptcy of US investment bank Lehman Brothers on 15 September 2008, showed how interconnected banks were (and are). This was not the beginning of the crisis—we have seen that the contraction of aggregate demand in the US began with the troubles in the housing market—but it signalled its escalation at the national and global level.

Thus the banking system, like an electricity grid, is a network. The failure of one of the elements in this connected network—whether a household or another bank—creates pressure on every other element. Just as happens in an electricity grid, the process in a banking system may create a cascade of subsequent failures, as occurred between 2006 and 2008. In our *Economist in action* video, Joseph Stiglitz, one of the

few economists who warned repeatedly about the risks inherent in the financial system in the lead-up to the financial crisis, explains the combination of incentives, external effects and positive feedback processes that led to this cascade of financial failure.

DISCUSS 17.6: BEHAVIOUR IN THE FINANCIAL CRISIS

Watch this explanation of the behaviour of households and banks in the financial crisis.

1. Which models that you have used in this unit can you fit to the story told in the video?
2. Are there parts of the video that you cannot explain using this unit?

DISCUSS 17.7: POST-CRISIS POLICIES

1. Compare the Great Depression and the 2008 financial crisis in terms of their institutional context (for example rules about exchange rates, trade union power).
2. How did these institutions affect the success of the policy responses by governments?

17.11 CONCLUSION

The human body is a miracle of self-stabilising processes, mobilising an immune system to repel infection and restoring damaged tissue or even brain cells. The *homeostatic* mechanisms of a modern economy are often similarly miraculous.

What have economists learned in the last century? One hundred years ago economists would have thought that the economy is always reliably self-correcting. Now they understand it is more like the human body: sometimes the economy's homeostatic mechanisms are overwhelmed, and it needs a doctor.

Economists learned the lesson of the Great Depression. They learned that an economy could get stuck with low output and high unemployment because of a lack of adequate aggregate demand. And they learned that the self-stabilising behaviour of private sector actors couldn't be relied on to end the crisis because of positive feedback processes. As we have seen, new policy regimes were developed at the national and international level after the second world war. Governments introduced or broadened the scope of stabilising mechanisms like unemployment insurance.

Economists learned about the importance of aggregate demand, but it gave them an undue confidence that a combination of fiscal and monetary policy would virtually eliminate unemployment in the long run. This helps explain most economists' failure to diagnose the supply-side character of the first oil shock in 1973. Figure 17.25 illustrates this policy mistake for the US. The doubling of the oil price (in real terms) is indicated by the increase in the index from 5 to 10 in the chart in 1973. From Unit 14 and this unit, we know that when the national economic pie is reduced by a commodity price shock, this will intensify the conflict of interest over its division, and so inflation increased to more than 10% in 1974. Yet policymakers were focused on the effect of the oil price shock in reducing aggregate demand and raising unemployment. They responded by loosening monetary policy (look at the falling nominal and real interest rates). Fiscal policy was not tightened.

A different response followed the second oil shock in 1979. The focus was on the need to reduce inflation and restore expected profits. Economists shifted their attention away from aggregate demand to the supply side of the economy. Policymakers used supply-side policies closely associated with Prime Minister Margaret Thatcher in the UK and President Ronald Reagan in the US.

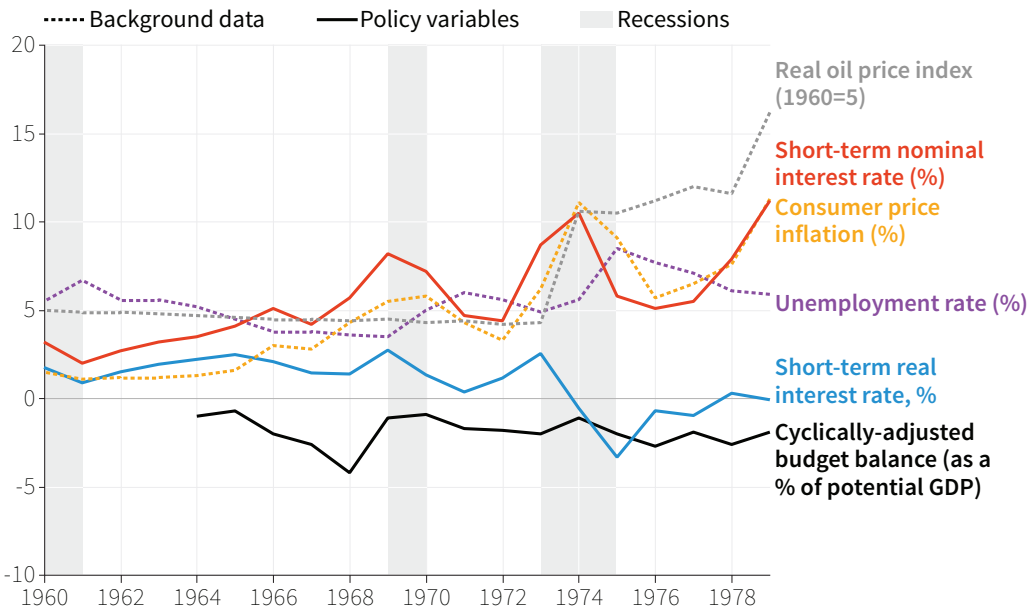


Figure 17.25 *Policy choices during the end of the golden age: United States (1960-79).*

Source: *Federal Reserve Bank of St Louis (FRED); Congressional Budget Office; US Bureau of Labor Statistics.*

In Figure 17.26 we summarise the lessons for economists from each epoch.

EPOCH	DATES	PRIOR CONVENTIONAL WISDOM	THE LESSON	WHAT ECONOMISTS LEARNED	PRIMARY AUTHOR
1920s AND GREAT DEPRESSION	1921-1941	Markets are self-correcting, efficient, and ensure the full use of resources.	Collapse of aggregate demand, high and persistent unemployment.	• Instability is an intrinsic feature of the aggregate economy • Aggregate demand can be stabilised by government policy • Demand matters	Keynes
GOLDEN AGE OF CAPITALISM AND ITS DEMISE	1948-1979	Government policy can implement an employment target by picking a point on the Philips curve.	Late 60s decline in profits, investment and productivity growth. Stable Phillips curve trade-off disappears.	• With given institutions, the need to maintain profits, investment and productivity growth can limit the ability of a government to implement sustainable low unemployment • Supply matters • Institutions matter	Friedman
FROM STAGNATION TO THE FINANCIAL CRISIS	1979-2013	Instability has been purged from capitalist dynamics; minimally regulated financial markets work well.	Financial and housing market crash of 2008.	• Debt-fuelled financial and housing bubbles will destabilise an economy in the absence of appropriate regulations • Institutions matter • Money matters	Minsky

Figure 17.26 *The economy as teacher: What economists learned in the three epochs.*

We can draw three conclusions:

1. *Economists have learned from the successes and the failures of the three epochs:* Though the process has been slow, economics today is the result of this process.
2. *Successful policies in each epoch did not prevent positive feedback processes that contributed to subsequent crises:* Each epoch succeeded initially because the policies and institutions that had been adopted addressed the shortcomings of the previous epoch. But then policymakers and economists have been taken by surprise when virtuous circles have turned into vicious circles.

3. *No school of thought has policy advice that would have been good in every epoch:* The value of competing approaches and insights depends on the situation. Ideas from both Friedman and Keynes have been essential to what economists have learned.

When Germany invaded France in 1914 at the beginning of the first world war, the French soldier Andre Maginot was wounded in the attack. When he later became minister of war he was determined to construct an impregnable line of defence, which we remember as *The Maginot Line*, in case German soldiers tried to march into France again.

At the beginning of the second world war Germany's *blitzkrieg* (lightning war) attack used tanks and motorised troop carriers. They didn't breach the Maginot line. They didn't have to: they drove around it instead.

Economists today are trying to avoid Maginot's error. A careful study of the economic history of the past century will help us not always to fight the "last war", and to prepare for whatever new difficulties will arise.

CONCEPTS INTRODUCED IN UNIT 17

Before you move on, review these definitions:

- *Positive feedbacks*
- *Global financial crisis*
- *Golden age of capitalism*
- *Great Depression*
- *Gold standard*
- *Catch-up growth*
- *Oil price shocks*
- *Subprime mortgage*
- *Stagflation*
- *Effective tax rate on profits*
- *Postwar accord*
- *Financial accelerator*
- *Financial deregulation*
- *Great moderation*
- *Great recession*
- *Zero lower bound*
- *Bank bailouts*
- *Austerity policy*

DISCUSS 17.8: HOOVER'S BALANCED BUDGET

On 4 April 1932, as the US economy spiralled downward, President Hoover wrote to the US Congress (you can read the letter here) to advocate a balanced budget and cuts in government spending.

Write a critique of Hoover's letter, using the economics in Units 11 to 17.

DISCUSS 17.9: AUSTERITY POLICY

In Unit 13 we introduced the *paradox of thrift* and examined the use of *austerity policies* in many countries before their economies had recovered from the recession that followed the 2008 crisis.

Were the lessons of the Great Depression forgotten when austerity policies were introduced? (This analysis written by Barry Eichengreen and Kevin O'Rourke will help you.)

Key points in Unit 17

Positive feedback

This can turn what would otherwise have been an ordinary downturn into a major fall in output, as occurred following the stock market crash of 1929 and the financial crisis of 2008.

The golden age

During this epoch institutional change and positive feedbacks supported rapid growth in investment, productivity and wages; but the result was unsustainable due to negative effects on profits, leading to the period of stagflation.

The role of household wealth

We can understand the dynamics of the three epochs by tracking the wealth of households and their attempts to adjust to the shocks of unemployment or falling house prices.

The roles of demand, wages and finance

The three epochs taught economists that aggregate demand matters; that successfully pursuing high employment involves detaching the wage bargain from the unemployment rate; and that unregulated financial markets are prone to instability.

Distinct actors invest, work and save

Their interests are sometimes in conflict, and private contracts or government policies cannot adequately regulate their actions. This is the source of dynamism, but also of instability in the capitalist economy.

17.12 READ MORE

Bibliography

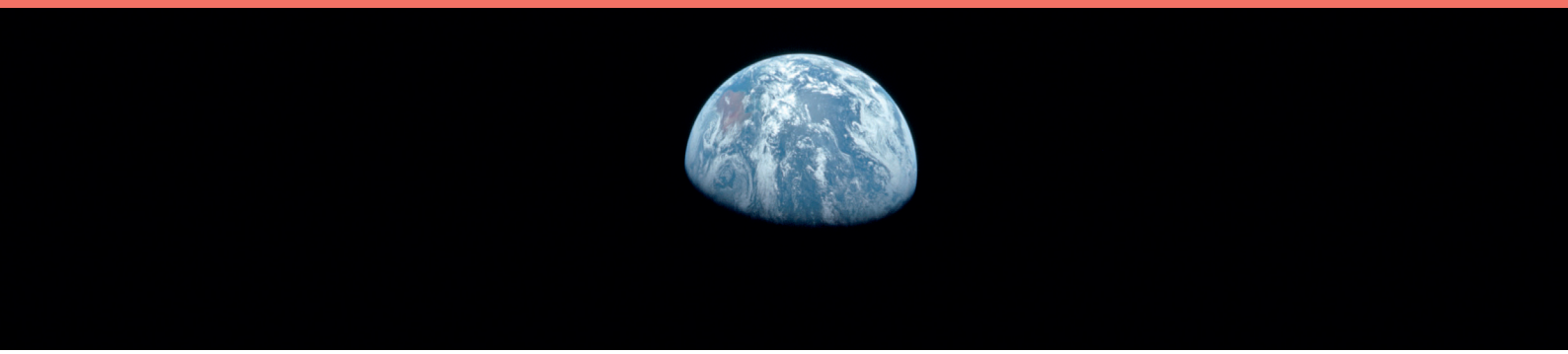
1. Almunia, Miguel, Agustín Bénétrix, Barry Eichengreen, Kevin H. O'Rourke, and Gisela Rua. 2010. 'From Great Depression to Great Credit Crisis: Similarities, Differences and Lessons.' *Economic Policy* 25 (62): 219–65.
2. Alvaredo, Facundo, Anthony B Atkinson, Thomas Piketty, Emmanuel Saez, and Gabriel Zucman. 2016. 'The World Wealth and Income Database (WID).'
3. Ball, Philip. 2002. 'Blackouts Inherent in Power Grid.' *Nature News*, November.
4. Ball, Philip. 2004. 'Power Blackouts Likely.' *Nature News*, January.
5. Bank for International Settlements. 2015. 'Residential Property Price Statistics.' November 20.
6. Bank of England. 2012. *Financial Stability Report, Issue 31*.
7. Bean, Charles, and Nicholas Crafts. 1996. 'British Economic Growth since 1945: Relative Economic Decline... and Renaissance?' In *Economic Growth in Europe since 1945*, edited by Nicholas Crafts and Gianni Toniolo. Cambridge: Cambridge University Press.
8. Bernanke, Ben. 1983. 'Nonmonetary Effects of the Financial Crisis in the Propagation of the Great Depression.' *American Economic Review* 73 (3): 257–76.
9. Boltho, Andrea. 1989. 'Did Policy Activism Work?' *European Economic Review* 33 (9): 1709–26.
10. Bowles, Samuel, David M Gordon, and Thomas E Weisskopf. 1989. 'Business Ascendancy and Economic Impasse: A Structural Retrospective on Conservative Economics, 1979–87.' *Journal of Economic Perspectives* 3 (1): 107–34.
11. CPB Netherlands Bureau for Economic Policy Analysis. 2015. 'World Trade Monitor.'
12. Card, David, and Richard B. Freeman. 2004. 'What Have Two Decades of British Economic Reform Delivered?' In *Seeking a Premier Economy: The Economic Effects of British Economic Reforms, 1980 - 2000*, edited by David Card, Richard Blundell, and Richard B. Freeman. Chicago, Il: University of Chicago Press.
13. Carlin, Wendy, and David Soskice. 2014. *Macroeconomics: Institutions, Instability, and the Financial System*. Oxford: Oxford University Press.
14. Crafts, Nicholas, and Peter Fearon. 2013. 'Depression and Recovery in the 1930s: An Overview.' In *The Great Depression of the 1930s: Lessons for Today*, edited by Nicholas Crafts and Peter Fearon. Oxford: Oxford University Press.
15. Crockett, Andrew. 2000. 'Marrying the Micro- and Macro-Prudential Dimensions of Financial Stability.' September 21.
16. Eichengreen, Barry. 1996. 'Institutions and Economic Growth: Europe after World War II.' In *Economic Growth in Europe since 1945*, edited by Nicholas Crafts and Gianni Toniolo. Cambridge: Cambridge University Press.

17. Eichengreen, Barry. 2006. *The European Economy since 1945: Coordinated Capitalism and beyond*. Princeton, NJ: Princeton University Press.
18. Eichengreen, Barry, and Kevin O'Rourke. 2010. 'What Do the New Data Tell Us?' *VoxEU.org*. March 8.
19. Eichengreen, Barry, and Peter Temin. 2010. 'Fetters of Gold and Paper.' *Oxford Review of Economic Policy* 26 (3): 370–84.
20. Field, Alexander J. 2003. 'The Most Technologically Progressive Decade of the Century.' *American Economic Review* 93 (4): 1399–1413.
21. Fishback, Price. 2013. 'US Monetary and Fiscal Policy in the 1930s.' In *The Great Depression of the 1930s: Lessons for Today*, edited by Nicholas Crafts and Peter Fearon. Oxford: Oxford University Press.
22. Friedman, Milton, and Anna Jacobson J. Schwartz. 1982. *Monetary Trends in the United States and the United Kingdom, Their Relation to Income, Prices, and Interest Rates, 1867-1975*. Chicago, Il: University of Chicago Press.
23. Glyn, Andrew. 2006. *Capitalism Unleashed: Finance, Globalization, and Welfare*. Oxford: Oxford University Press.
24. Glyn, Andrew, Alan Hughes, Alain Lipietz, and Ajit Singh. 1989. 'The Rise and Fall of the Golden Age.' In *The Golden Age of Capitalism: Reinterpreting the Postwar Experience*, edited by Stephen A. Marglin and Juliet Schor. New York, NY: Oxford University Press.
25. Gordon, Robert J. 1986. *The American Business Cycle: Continuity and Change*. Chicago, Il: University of Chicago Press.
26. International Monetary Fund. 2009. *World Economic Outlook: January 2009*. IMF.
27. International Monetary Fund. 2013. 'IMF Fiscal Monitor April 2013: Fiscal Adjustment in an Uncertain World, April 2013.' April 16.
28. Johnson, Simon, and James Kwak. 2010. *13 Bankers: The Wall Street Takeover and the next Financial Meltdown*. New York, NY: Knopf Doubleday Publishing Group.
29. Krenn, Robert, and Robert J Gordon. 2010. 'The End of the Great Depression 1939-41: Policy Contributions and Fiscal Multipliers.' *NBER Working Papers 16380*, September.
30. Lanchester, John. 2011. 'How We Were All Misled.' *The New York Review of Books*. December 8.
31. Mayer, Gerald. 2004. *Union Membership Trends in the United States*. Washington, DC: Congressional Research Service.
32. Mian, Atif, and Amir Sufi. 2014. *House of Debt: How They (and You) Caused the Great Recession, and How We Can Prevent It from Happening Again*. Chicago, Il: The University of Chicago Press.
33. Minsky, Hyman P. 1975. *John Maynard Keynes*. New York, NY: McGraw-Hill.
34. Minsky, Hyman P. 1982. *Can 'It' Happen Again? Essays on Instability and Finance*. Armonk, NY: M.E. Sharpe.
35. OECD. 2015. 'OECD Statistics.'
36. Olney, Martha. 1999. 'Avoiding Default: The Role of Credit in the Consumption Collapse of 1930.' *The Quarterly Journal of Economics* 114 (1): 319–35.

37. *Oxford Review of Economic Policy*. 2010. Lessons from the 1930s, 26 (3).
38. Ramey, Valerie A. 2011. 'Can Government Purchases Stimulate the Economy?' *Journal of Economic Literature* 49 (3): 673–85.
39. Reinhart, Carmen M, and Kenneth S Rogoff. 2009. *This Time Is Different: Eight Centuries of Financial Folly*. Princeton, NJ: Princeton University Press.
40. Romer, Christina D. 1990. 'The Great Crash and the Onset of the Great Depression.' *The Quarterly Journal of Economics* 105 (3): 597–624.
41. Romer, Christina D. 1992. 'What Ended the Great Depression?' *The Journal of Economic History* 52 (04): 757–84.
42. Santa Fe Institute. 2011. 'Forest Fire Mathematics Suggests Less Fire Suppression.' *Physical Review E*. November 16.
43. Shin, Hyun Song. 2009. 'Discussion of "The Leverage Cycle" by John Geanakoplos'.
44. Temin, Peter, and Barrie A Wigmore. 1990. 'The End of One Big Deflation.' *Explorations in Economic History* 27 (4): 483–502.
45. The Conference Board. 2014. 'Total Economy Database.'
46. *The Economist*. 2012. '1929-33: The Big One.' In *The Slumps That Shaped Modern Finance*, Part 6.
47. Toniolo, Gianni. 1998. 'Europe's Golden Age, 1950-1973: Speculations from a Long-Run Perspective.' *The Economic History Review* 51 (2): 252–67.
48. US Federal Reserve. 2015. 'Financial Accounts of the United States, Historical.' December 10.
49. United States Bureau of the Census. 2003. *Historical Statistics of the United States: Colonial Times to 1970, Part 1*. United States: United States Govt Printing Office.
50. Wallis, John Joseph. 2000. 'American Government Finance in the Long Run: 1790 to 1990.' *Journal of Economic Perspectives* 14 (1): 61–82.

18

ECONOMICS AND THE ENVIRONMENT



HOW THE PRODUCTION AND DISTRIBUTION OF GOODS AND SERVICES AFFECTS THE FRAGILE BIOSPHERE OF OUR PLANET, AND HOW THE RESULTING ENVIRONMENTAL PROBLEMS CAN BE ADDRESSED

- Production and distribution of goods and services unavoidably alter the biosphere
- Climate change resulting from economic activity is a major threat to future human wellbeing, and it illustrates many of the challenges of designing and implementing appropriate environmental policies
- Environmental policy should implement least-cost ways of abating environmental damages. In selecting the level of abatement it should balance the cost of reducing environmental damage against the opportunity costs of doing so
- Policies should be evaluated on the grounds of efficiency and fairness, taking account of the distribution of costs and benefits among different groups in a society, citizens of different countries, and people in future generations
- Some policies work by using taxes, subsidies or other policies to alter prices so that people internalise the external effects of their production and consumption decisions; other policies directly prohibit or limit the use of environmentally damaging materials and practices
- Environmental policies can act as a stimulus to “green” innovation
- Social preferences may make the implementation of environmental policies easier if economic actors (citizens, consumers and owners of firms) place a positive value on the environment, and on the wellbeing of others—including future generations

Institute for New Economic Thinking
UCL SciencesPo Azim Premji University
FRIENDS PROVIDENT Foundation

In 1980, one of the most famous bets in science history took place. Paul Ehrlich, a biologist, predicted that rapidly increasing population would make natural *resources* scarcer. Julian Simon, an economist, thought that humanity would never run out of anything because higher prices would stimulate the search for new *reserves*, and ways of economising on the use of resources. Ehrlich bet Simon that the price of a basket of five commodities—copper, chromium, nickel, tin, and tungsten—would increase in real terms over the decade, reflecting increased scarcity.

On 29 September 1980 they bought $200 of each of the five commodities (a total wager of $1,000). If prices of these resources went up faster than inflation over the next 10 years, Simon would pay Ehrlich the difference between the *inflation-adjusted prices* and $1,000. If real prices fell, Ehrlich would pay Simon the difference. During that time, the global population increased by 846 million (19%). Also during that time, income per person increased by $753 (15%, adjusted for inflation in 2005 dollars). Yet, in those 10 years, the inflation-adjusted prices of the commodities fell from $1,000 to $423.93. Ehrlich lost the bet and sent Simon a cheque for $576.07.

The Ehrlich-Simon bet was motivated by the question of whether the world was "running out" of natural resources, but an interval of 10 years is unlikely to tell us much about the long-run scarcity of raw materials. The basic framework of supply and demand (see Units 8 and 9) tells us why. Commodities such as copper or chromium generally have inelastic (steep) short-run demand and supply curves, because there are few substitutes for these resources. This means that relatively small demand and supply shocks generate large and sudden changes in the market-clearing price.

The market for crude oil clearly demonstrates this. Figure 18.1a plots, for 1861 to 2014, the real price of oil in world markets in constant 2014 US dollars and, from 1965, the total quantity consumed globally in million barrels per day. The price of oil shows large fluctuations, but the path of world oil consumption is much smoother. To understand what drives these fluctuations, we need to use the supply and demand model.

Figure 18.1b shows an index of global commodity prices since 1960. You can easily see the effect of oil price shocks in the 1970s and 2000s. In the 1970s, supply disruptions were responsible for a leftward shift of the supply curve. The 2000s was a period of rapid economic growth in industrialising countries, especially China and India. The result was a shift to the right of the demand curve. When global growth slowed sharply with the crisis of 2008-9, the demand curve shifted left.

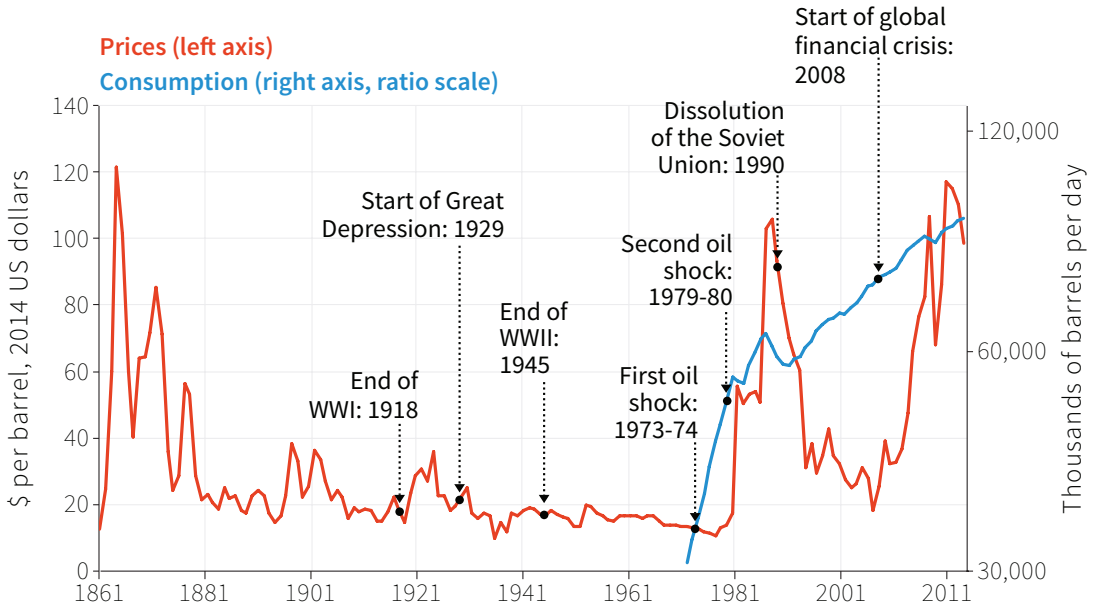


Figure 18.1a *World oil price in constant prices (1865-2014) and global oil consumption (1965-2014).*

Source: *BP Global. 2015. 'Statistical Review: Energy Economics.'*

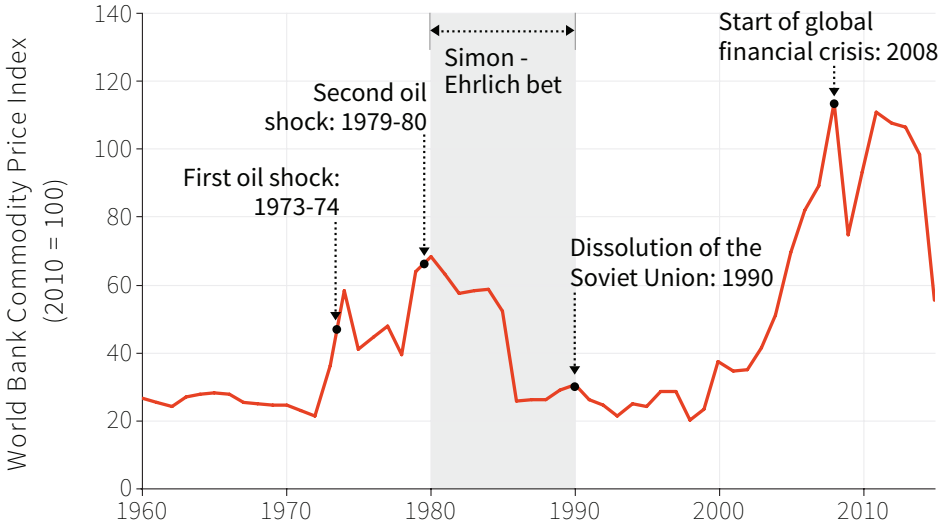


Figure 18.1b *Global commodity prices (1960-2014).*

Source: *The World Bank. 2015. 'Commodity Price Data.'*

What the two were really betting on was the race between two influences on commodity prices:

- *Ehrlich*: Increases in demand due to population growth and growing affluence would outstrip supply.
- *Simon*: The discovery of technologies to find new resources and extract them more efficiently would outstrip increases in demand.

From the year of the bet in 1981 until 2014, world reserves of oil more than doubled to 1.7 trillion barrels—in spite of the fact that more than 1 trillion barrels was extracted and consumed over those years.

At the time of the Ehrlich-Simon wager, some people were more concerned with the impact of population and economic growth on habitat destruction and *biodiversity loss*, pollution, degradation of environmental amenities and global climate change than on the price of nickel. You already know from Unit 1 that, had the bet been placed on whether the world would be warmer in 1990 than in 1980, Ehrlich would have won. And he would have won a similar bet had he struck it in most of the decades since 1850.

The transformation of living standards since the Industrial Revolution has been possible because of the combination of human ingenuity and available resources in the form of air, water, soil, metals, hydrocarbons like coal and oil, fish stocks and so on. These were all once abundant and, apart from the costs of extraction, they were free. Some, like hydrocarbons, are still abundant (look ahead to Figure 18.6); others, like unpolluted air and water, are becoming scarce.

In some cases the fragility of our environment under pressure from the growth of economic activity can lead not only to progressive degradation, but also to accelerating, self-reinforcing collapse. An example is the Grand Banks cod fishery, in the north of the Atlantic Ocean. In the 18th and 19th centuries, legendary schooners such as the Bluenose (Figure 18.2) raced back to port to sell their catch to be the first on the market, and to offer fresh fish. By the late 20th century, the Grand Banks had sustained the livelihoods of US and Canadian fishing communities for 300 years.

Figure 18.2 *The Grand Banks fishing schooner, The Bluenose.*

Then, suddenly, the fishing industry in the Grand Banks died, as did many of the old fishing towns. Figure 18.3 gives the quantity of cod caught over 163 years, showing a gradual upward trend and a pronounced spike coinciding with the introduction of industrial fishing less than 50 years before the eventual disappearance of cod from the Grand Banks. We do not know if the cod will come back in their previous numbers in the Atlantic, although North Sea fisheries are now recovering after governments imposed restrictions on fishing.

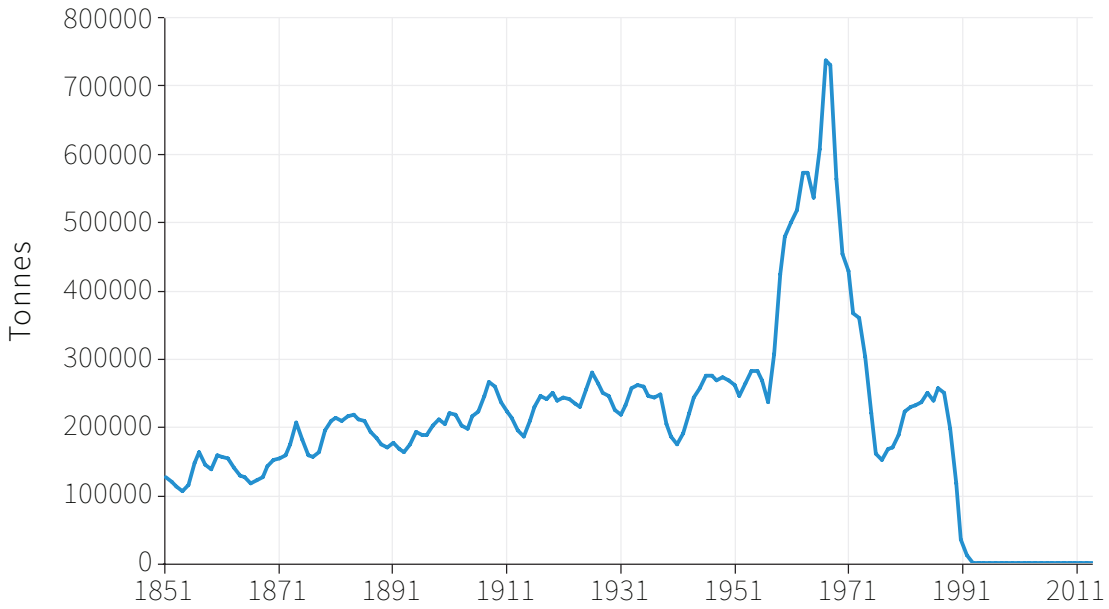


Figure 18.3 *The Grand Banks (North Atlantic) fisheries: Cod landings in tons (1851-2014).*

Source: *Millennium Ecosystem Assessment. 2005. Ecosystems and Human Well-Being: Synthesis. Washington, DC: Island Press.*

Ecosystem collapse hasn't happened only in the Grand Banks. We hear about the "death" of lakes, or the threat to the Amazon rainforest as a result of the deforestation from the expansion of farming, for example. These cataclysmic and rapid changes are an environmental vicious circle. In the Amazon, for example, change may become self-reinforcing:

- Farming reduces forest area.
- Deforestation reduces rainfall.
- Drought conditions increase the likelihood of fires.
- The forest dies back further, eventually passing a *tipping point*.
- Cumulative, self-reinforcing deforestation occurs independent of any further expansion of farming.

Similarly, the process of global warming can be self-reinforcing due to its impact on Arctic ice cover:

- Warming reduces the extent of sea ice cover.
- Open water reflects less solar radiation than sea ice.
- This is an additional contribution to global warming.
- It further reduces the extent of sea ice cover.

Ecologists concerned about the impact of a growing economy on the planet sometimes liken our situation to that of a pond being taken over by a pondweed that would kill everything else in the water (and, ultimately, the pondweed itself). Suppose that each morning there's twice as much pondweed as there was the day before, and we know that in 30 days the pond would be choked with the weed if we didn't do anything.

ENVIRONMENTAL TIPPING POINT

- On one side of a tipping point, processes of environmental degradation are self-limiting.
- On the other side, positive feedbacks lead to self-reinforcing, runaway environmental degradation.

But say we preferred to wait until the pond was half-choked with weed until we did anything about it. How much time would we have to act? When would the pond be half full of weeds?

On the 29th day.

We would have a single day to save the pond.

To many ecologists, the moral of this story is that time is running out. If we act like pondweed, the planet (our pond) cannot possibly sustain our increasing production and consumption of resources.

But as James Boyce, an environmental economist, also points out, we are not pondweed:

"Each pondweed organism is pretty much like any other. But humans differ greatly from one another, both in their impacts on the environment and in their ability to shield themselves from these impacts."
James K. Boyce, *Economics, the Environment and Our Common Wealth (2012)*

We differ from pondweed (and other nonhuman organisms) because we can reason about the merits of possible remedies to abate the impacts we have on our environment, and because we have the potential to adopt policies to address these problems.

DISCUSS 18.1: SELF-REINFORCING PROCESSES

Self-reinforcing processes such as the ones described above do not just happen in nature. In Unit 17, for example, we discussed how increases in house prices can reinforce a boom and become self-sustaining.

Explain in what ways the cumulative self-reinforcing processes described by environmental scientists are similar to (or different from) processes that occur in a housing or stock price bubble.

18.1 EXTERNAL EFFECTS, INCOMPLETE CONTRACTS AND MISSING MARKETS

In Unit 1, we saw that the production and distribution of goods and services—economic activity—takes place within the biological and physical system. In this unit we investigate the nature of the global ecosystem that sustains us by providing the resources that feed economic processes, and also the sinks where we dispose our wastes. As we saw in Figure 1.8 and Figure 1.18, the economy is embedded within our society, but also within the ecosystem. Resources (matter and energy) flow from nature into the human economy. Waste, such as carbon dioxide (CO_2) emissions, or toxic sewage produced by firms and households, flows back into nature—mainly into the atmosphere and the ocean. Scientific evidence suggests that the planet has a limited capacity to absorb the pollutants that the human economy generates.

In Unit 4 we introduced environmental problems at a local level among people who were similar in most respects. Anil and Bala were neighbouring landowners with a pest management problem. They could choose between an environmentally damaging pesticide and a benign pest management system. The outcome was inefficient—and environmentally destructive—because they could not make a binding agreement (a complete and enforceable contract) about how they would act in advance. In Unit 4 we also discovered that contributing to sustaining the quality of the environment is, to some extent, a public good, and that there are strong self-interested motives to free ride on the activities of others. So, while everyone would benefit if we all contributed to protecting the environment, we often do not.

However, when just a few individuals interact, we saw that informal agreements and social norms (a concern for the others' wellbeing, for example) might be sufficient to address environmental problems. Examples found in real life included irrigation systems and the management of common land.

In Unit 10 we expanded the scope of environmental problems to include two classes of people pursuing different livelihoods. We considered a hypothetical pesticide called Weevokil (based, again, on real-world cases) and its effects on fishing and the jobs of workers who produce bananas. In this case markets were missing—the plantation owners did not buy the right to pollute the fisheries. They could do it for free. This is just another case of an incomplete contract.

In cases like this, taxes that increase the polluter's marginal private cost of production so that it equals the marginal social cost achieve an efficient reduction in production (and pollution). In this case solutions to the environmental problems—the external effects of the pesticide on the downstream fisheries—included bargaining between the organisations of fishermen and the plantation owners, and legislation. (In the real world case that inspired our Weevokil model, the government eventually banned the chemical).

The segment of Figure 10.11 that we reproduce in Figure 18.4 summarises the nature of market failures in interactions of economic actors with the environment, and some possible remedies.

THE DECISION	HOW IT AFFECTS OTHERS	COST OR BENEFIT	MARKET FAILURE (MISALLOCATION OF RESOURCES)	POSSIBLE REMEDIES	TERMS APPLIED TO THIS TYPE OF MARKET FAILURE
A firm uses a pesticide that runs off into waterways	Downstream damage	Private benefit, external cost	Overuse of pesticide and overproduction of crop in which it is used	Taxes, quotas, bans, bargaining, common ownership of all affected assets	Negative external effect, environmental spillovers *(Section 10.1)*
You take an international flight	Increase in global carbon emissions	Private benefit, external cost	Overuse of air travel	Taxes, quotas	Public bad, negative external effect *(Section 10.5)*

Figure 18.4 *External environmental effects.*

In this unit we consider the problem of climate change. Returning to the wager between Simon and Ehrlich, we can see that if they wanted to bet on climate change instead of mineral resources, there's immediately a problem: they could not have bet on a price. Climate does not have a price. Climate change is a problem of a

missing market that is global in scope. It involves people with vastly differing interests, ranging from those whose entire nation may be submerged by rising sea levels to those who profit from the production and use of carbon-based energy that contributes to global climate change. We will see that many of the concepts developed already—feasible sets and indifference curves—apply in these cases as well. But some new concepts will be necessary.

We move from asking why environmental problems arise to studying what might be done about them. To begin, we take the same approach to this problem as we did when we asked how Alexei the student or Angela the farmer decides how many hours to study or to work, or how the firm decides what price to set. In all cases we want to do the best we can when facing trade-offs between competing objectives.

First we ask, given that environmental quality is one among many goods that people prefer and that having more of one may require having less of another, how do we decide what mix of environmental quality and the other goods we would like to have? In later sections we consider conflicts of interest when we determine the level of environmental quality, and the policies that we might adopt to reach that goal.

18.2 CLIMATE CHANGE

From the US atom bomb attacks on Hiroshima and Nagasaki at the end of the second world war, until the end to the Cold War half a century later, nuclear holocaust was the *Armageddon*—the nightmare of total destruction—that haunted humanity.

Today, cataclysmic climate disruptions due to global warming are a similar nightmare. Like nuclear war, an Armageddon of climate change remains unlikely. But it cannot be ruled out; and many scientists now see climate change as the greatest threat to human wellbeing in our future.

Climate change is not the only serious environmental problem. Others include:

- The loss of biodiversity through species extinctions
- Lack of access to clean water
- The limits of the waste-carrying capacity of the globe's oceans
- Loss of natural assets due to desertification, deforestation, degradation of fresh water bodies (through chemical runoffs) and other processes

We focus on climate change because of its importance as a problem, and because it illustrates the difficulties of designing and implementing adequate environmental policies. This problem tests our framework of efficiency and fairness to the limit, because of four distinctive features:

- *Capping emissions is not sufficient:* The science of climate change indicates that the external effects of *greenhouse gas* emissions arise from the accumulation of carbon and other greenhouse gases in the atmosphere rather than from the annual flow of emissions. Stabilising emissions at current levels will not be enough, because the stock of greenhouses gases would continue to increase.
- *The worst-case scenario:* Experts are uncertain about the scale, timing and global pattern of the effects of climate change, but few rule out the small chance of a catastrophic and or irreversible outcome. Therefore a best guess or average of the scientific forecasts linking the concentration of greenhouse gases, global temperature and its effects should not be the only guide to policy.
- *A global problem requiring cooperation:* The contributions to climate change come from all parts of the world, and its effects will be felt by almost 200 autonomous nations. It will be solved only by unprecedented cooperation among at least the largest and most powerful nations.
- *Conflicts of interest:* The impacts of climate change differ across the globe and arise from different past activities. Future generations will experience the effects of today's emissions, and the actions we take to reduce them. How should we think about the costs it is fair to bear today, to take account of the lives and needs of total strangers from entirely different cultures and future generations?

Climate change and economic activity

The last 250 years of the 100-year climate hockey stick in Figure 18.5 reminds us of the connection between the industrial revolution and the concentration of carbon in the atmosphere. Figure 18.5 shows the data on the stock of CO_2 (in parts per million) using the right-hand scale, and global temperature (as the deviation from the average over the period 1961-1990) using the left-hand scale, for the period since 1750.

Burning fossil fuels for power generation and industrial use leads to emissions of CO_2 into the atmosphere. These activities, with CO_2 emissions from land-use changes, generate greenhouse gases equivalent to around 36 billion tonnes of CO_2 each year. Concentrations of CO_2 in the atmosphere have increased from 280 parts per million in 1800 to 400 parts per million, currently rising at 2-3 parts per million each year. CO_2 allows incoming sunlight to pass through it, but traps reflected heat on Earth, leading to increases in atmospheric temperatures and changes in climate. Some CO_2 also gets absorbed into the oceans. This increases the acidity of the oceans, killing marine life.

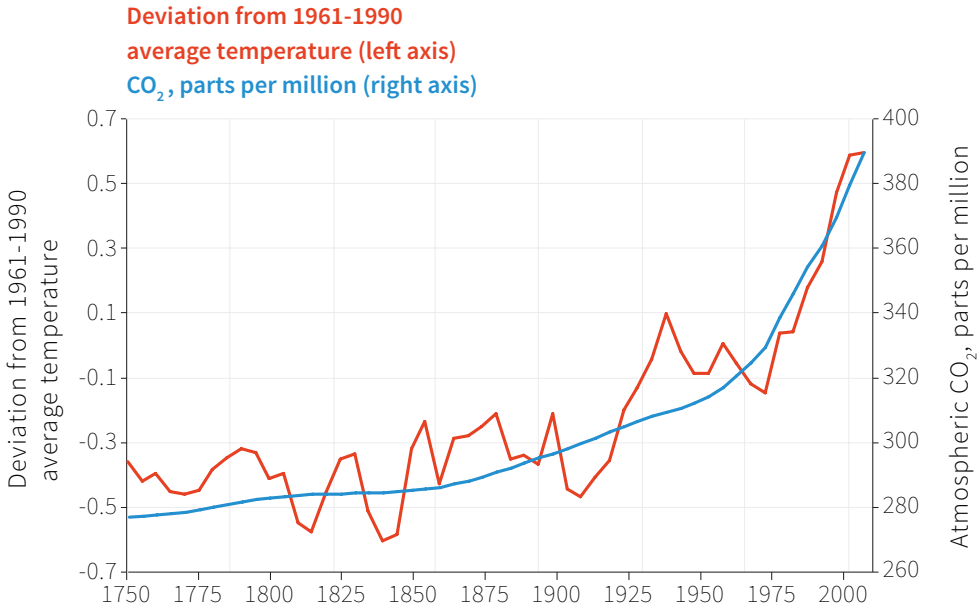


Figure 18.5 *Global atmospheric concentration of carbon dioxide and global temperature (1750-2010).*

Source: *Years 1010-1975: Etheridge, D. E., L. P. Steele, R. J. Francey, and R. L. Langenfelds. 2012. 'Historical Record from the Law Dome DE08, DE08-2, and DSS Ice Cores.' Division of Atmospheric Research, CSIRO, Aspendale, Victoria, Australia. Years 1976-2010: Data from Mauna Loa observatory. Boden, T. A., G. Marland, and R. J. Andres. 2010. 'Global, Regional and National Fossil-Fuel CO2 Emissions.' Carbon Dioxide Information Analysis Center (CDIAC) Datasets. Note: This data is the same as in Figures 1.7a and 1.7b. Temperature is average Northern hemisphere temperature.*

We can emit only a further 1 to 1.5 trillion tonnes of CO_2 into the atmosphere to give reasonable odds of limiting the increase in temperature to 2C more than pre-industrial levels. Should we manage to achieve this limit on emissions, there is still a probability of around 1% that temperature increases would be more than 6C, causing a global economic catastrophe. If we exceed the limit and temperature rises to 3.4C above pre-industrial levels, the probability of a climate-induced economic catastrophe would rise to 10%.

Figure 18.6 shows the temperature increase arising from the CO_2 emitted, which would be generated at different levels of use of the fossil fuel *reserves* (which can be technologically and economically extracted) and *resources* (estimated total amounts) in the Earth's crust.

Figure 18.6 indicates that keeping the warming to 2C implies that the majority of fossil fuel reserves and resources would remain in the ground.

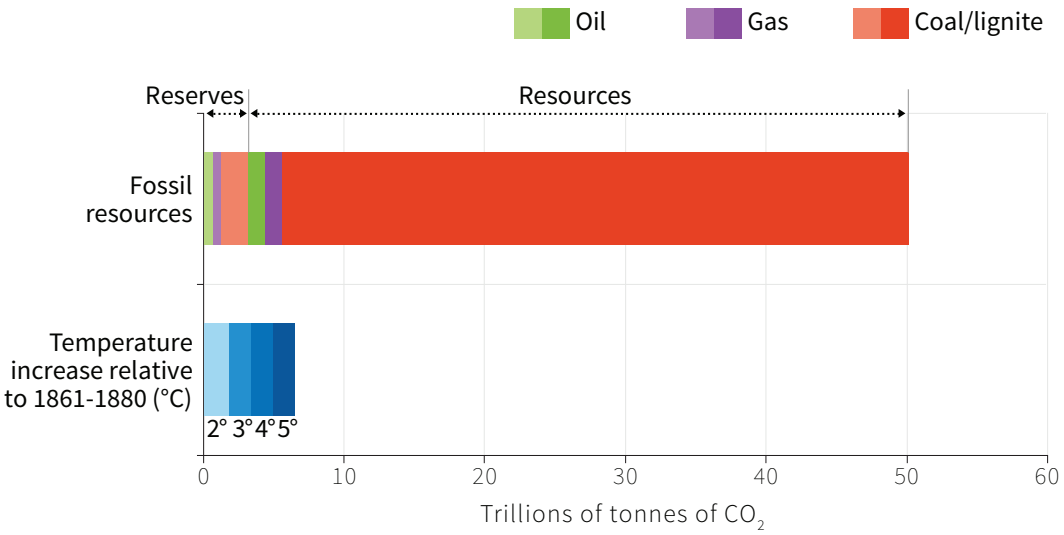


Figure 18.6 *Amount of carbon dioxide in fossil fuel reserves and resources exceeds the atmospheric capacity of the Earth as indicated by the extent of temperature increase.*

Source: *Calculations by Alexander Otto of the Environmental Change Institute, University of Oxford, based on: Aurora Energy Research. 2014. 'Carbon Content of Global Reserves and Resources'; Bundesanstalt für Geowissenschaften und Rohstoffe (The Federal Institute for Geosciences and Natural Resources). 2012. Energy Study 2012; IPCC. 2013. Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge: Cambridge University Press; Hepburn, Cameron, Eric Beinhocker, J. Doyne Farmer, and Alexander Teytelboym. 2014. 'Resilient and Inclusive Prosperity within Planetary Boundaries.' China & World Economy 22 (5): 76–92.*

DISCUSS 18.2: CLIMATE CHANGE CAUSES AND EVIDENCE

Use information that from the National Aeronautics and Space Administration web page on climate change, and the latest report of the Intergovernmental Panel on Climate Change to answer the following questions:

1. Explain what climate scientists believe to be the main causes of climate change.
2. What evidence is there to suggest that climate change is already occurring?
3. Name and explain three potential consequences of climate change in the future.
4. Discuss why the three consequences you have listed may lead to disagreements and conflicts of interest about climate policy.

18.3 THE ABATEMENT OF ENVIRONMENTAL DAMAGES

Climate change—like other environmental problems—can be addressed by environmental damage *abatement policies* such as:

- Discovering and adopting less-polluting technologies
- Choosing to consume fewer or less environmentally damaging goods
- Banning or limiting the use of environmentally harmful substances or activities

Policies may limit negative impacts on the environment by directly or indirectly inducing decision-makers to take account of the negative external effects that their choices impose on others. The cost of entirely eliminating the negative effects on the environment would surely exceed the benefits.

What environmental abatement policies should a nation adopt?

This is in part an economic question. It involves trade-offs between the goals of producing and consuming more, while enjoying a less degraded environment.

It is also an ethical question. It involves trade-offs between our consumption now and other people's environmental quality both now and in future generations. Therefore our policy choices raise questions not only of efficiency but also of fairness.

If we ask citizens about their views of the correct environmental policies, we expect their responses will differ because a deteriorating environment affects different people in different ways. Your point of view may depend on whether you work outdoors (you will benefit from a less polluted local environment) or in fossil fuel production (you may lose your job if the polluting firm shuts down as a result of higher abatement costs levied on the firm); it may depend on whether you have no choice but to live near a source of air pollution, or are wealthy enough to have a second home in the countryside.

Your opinion about how much we should spend today to protect *future* environments would no doubt differ from the values of those who make up the distant future generations that would be affected by our choices, if we could ask them. People's views are strongly influenced by their self-interest but, as you would expect from the behavioural experiments in Unit 4, not totally so. We worry about the effect on others, even total strangers.

For simplicity, we firstly set aside these differences and consider a population composed of identical individuals. We ignore future generations, or optimistically assume that we will all live forever. We will also assume that environmental quality

is a *pure public good*: everyone enjoys (or suffers) the same level of environmental quality. Later in this unit we will look at what changes when we do not make these assumptions.

As economists, how can we reason about the level of environmental quality that we would like to enjoy, knowing that people may have to consume less so they can enjoy a better environment? The first thing to think about is the actions that we can take and their consequences: the feasible set of outcomes.

To do this we need to consider the ways that the resources of the society could be diverted from their current uses to abate the environmentally degrading effects of economic activity. The nation may adopt abatement policies to limit environmental damage. Abatement policies include taxes on emissions of pollutants, and incentives to use fuel-efficient cars.

Abatement costs and the feasible set

To get some idea of how economists assess abatement policy options, we look at the cost of reduction of greenhouse gas emissions in Figure 18.7. The figure shows the relationship between potential abatement (measured in gigatonnes of CO_2 equivalent, a unit used to measure abatement by the International Panel on Climate Change), using specific changes in how economies across the globe work, and its cost per tonne. These estimates were made by the consultancy McKinsey. The science in this field is young, and technologies are continuously developing. As knowledge advances, the estimated abatement cost curve will change.

To interpret the data, note that for each method of reducing CO_2 emissions, a short bar means that there's a lot of abatement per dollar spent. A wider bar means that this method has a higher potential to abate emissions. A policymaker looks for short, wide bars.

We order the policies from the least abatement per dollar spent on the left to most abatement per dollar spent on the right. Policies to convert agriculture toward lower emissions are most efficient by this measure, through nuclear, wind, solar photovoltaic, and at the top retrofitting gas-fired power plants for carbon capture and storage, the highest-cost policy.

GLOBAL GREENHOUSE GAS ABATEMENT COST CURVE

This shows the total cost of abating *greenhouse gas* emissions using *abatement policies* ranked from the most cost-effective to the least.

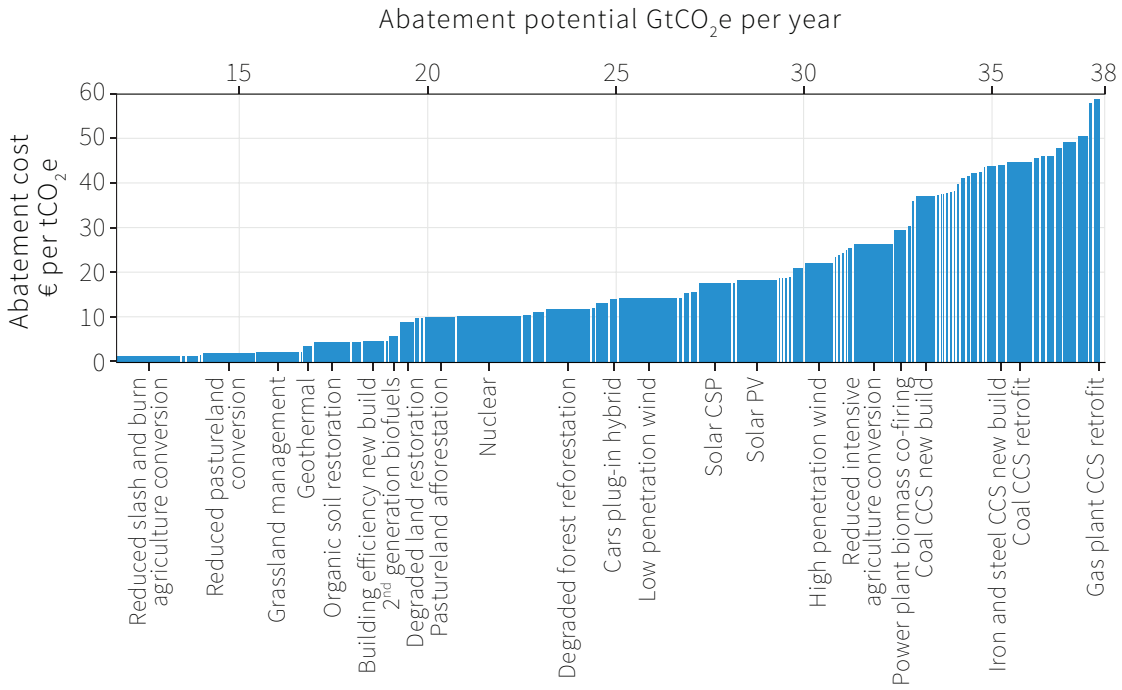


Figure 18.7 *Global greenhouse gas abatement curve: Abatement in 2030 compared with business as usual.*

Source: *McKinsey & Company. 2013. Pathways to a Low-Carbon Economy: Version 2 of the Global Greenhouse Gas Abatement Cost Curve. McKinsey & Company.*

But even focusing on only the most efficient bars, implementing abatement policies would divert resources from the production of other goods and services: the *opportunity cost* of an improved environment would be reduced consumption. (If you are wondering if this is always the case, look forward to section 18.9, and in particular, Figures 18.26 and 18.27).

We can use data like that in Figure 18.7 to estimate how much abatement we get for any level of expenditure, assuming we implement the most efficient methods first. These calculations give Figure 18.8. We would start by implementing the cheap and effective measures, such as land management and conversion policies. Having exhausted these policies, the curve becomes flatter at higher levels of expenditure, where we would be devoting resources to less efficient methods such as carbon capture and storage (CCS) modifications to power stations. See our Einstein section on marginal abatement costs and the total productivity of abatement expenditures for more detail on the calculations.

The curve in the figure is like a production function for abatement. It is a relationship between an input—in this case abatement expenditures—and an output—an improved environment. It is similar to the function describing Alexei's hours of study and the grade he gets, or Angela's work and the grain she produces.

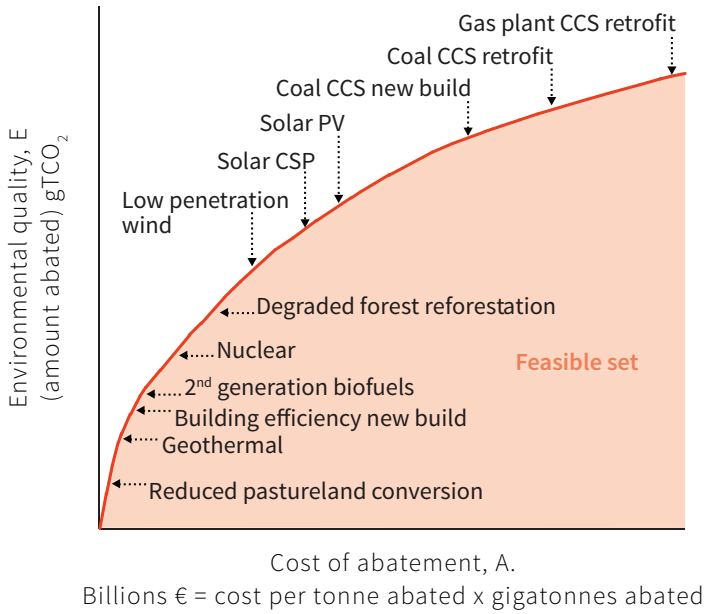


Figure 18.8 *The feasible set for climate change constructed from Figure 18.7: Abatement in 2030 compared with business as usual.*

Source: *McKinsey & Company. 2013. Pathways to a Low-Carbon Economy: Version 2 of the Global Greenhouse Gas Abatement Cost Curve. McKinsey & Company.*

Using figures like 18.8, we can establish all of the possible combinations of consumption and environmental quality that are feasible. The available abatement technology is shown by the shaded set of points in Figure 18.9. In this figure the horizontal axis measures the expenditure on abatement (for example, the cost per tonne of greenhouse gases abated, multiplied by the number of tonnes abated). The vertical axis measures environmental quality, or equivalently, abatement achieved. The zero point on the vertical axis is a situation in which zero abatement occurs.

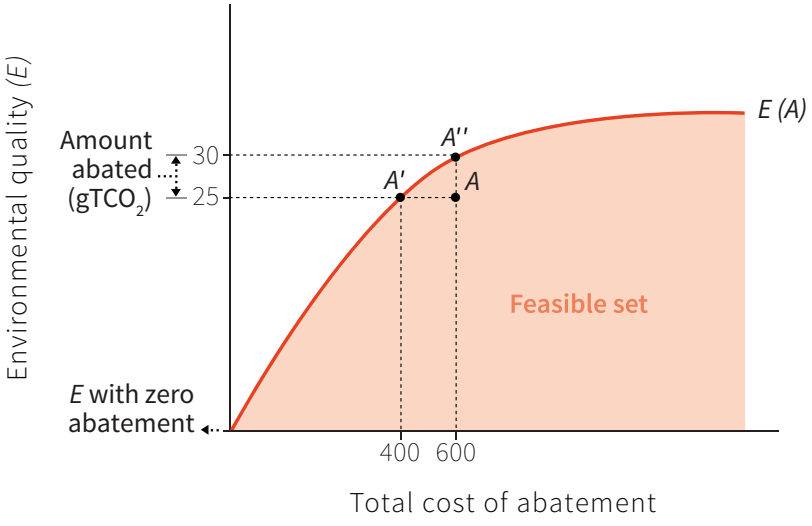


Figure 18.9 *The trade-off between consumption and environmental quality: Environmental quality rises as abatement costs are incurred (total cost of abatement is cost per tonne abated, multiplied by the number of tonnes abated).*

The shaded area is the feasible set of abatement expenditures and environmental outcomes. Points like *A* in the interior of the set are inefficient abatement policies. At *A*, we can see that there are alternative measures that would achieve the same level of abatement (25 gigatonnes) at lower cost (€400bn rather than €600bn). Similarly, for expenditure of €600bn, the choice of the most cost-effective abatement techniques would deliver 30 tonnes of CO_2 abatement and higher environmental quality than at point *A*. Economists say that a point like *A* is *dominated* by points *A'* and *A''* and all the points in between. This means that at any of these other points there could be lesser abatement costs and as much abatement (*A'*), or greater abatement at the same cost (*A''*).

How would an inefficient point like *A* in Figure 18.9 occur? In Figure 18.8 the policies were ordered so that the first expenditures on abatement are devoted to the most effective abatement policy. After exhausting the potential of each policy we moved to the next, less effective policy.

Figure 18.10 shows the abatement options based on the data in Figure 18.8, but with more costly policies adopted first. If a society has committed to spend €8.37bn on abatement, and spends it all on coal carbon capture, nuclear, and other less effective options, then the abatement cost curve would be as shown in Figure 18.10.

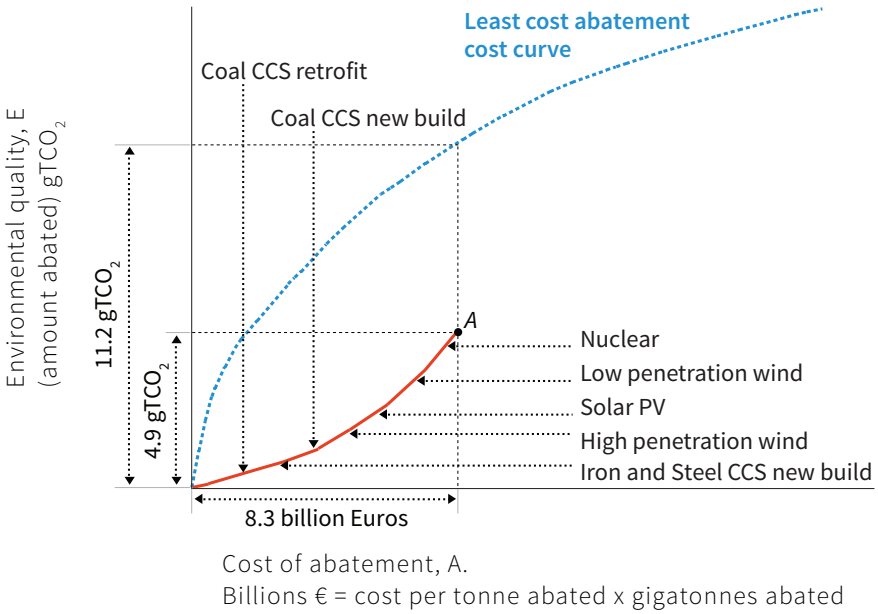


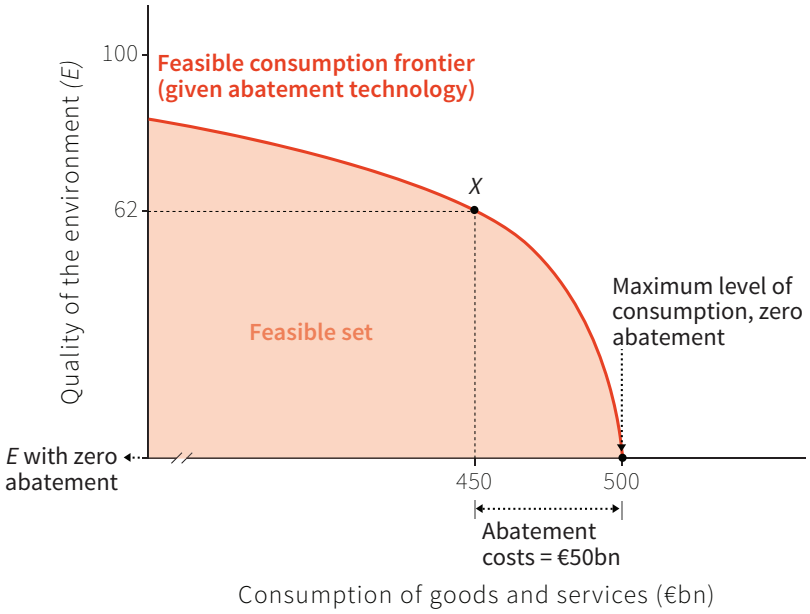
Figure 18.10 *An abatement cost curve in which more costly technologies are adopted first.*

We can see that if €8.37bn were spent on abatement, the level of abatement would be 4.94 gigatonnes of CO_2 not emitted, rather than the abatement of 11.2 gigatonnes that would have been possible had the society implemented least-cost policies, as shown in Figure 18.8.

Figures 18.8 and 18.10 send a clear message about priorities: if we have a limited amount to spend on abatement, it says, focus on agriculture. According to Figure 18.8, we should focus on nuclear power, solar and wind ahead of new coal plants or retrofitting old ones for carbon capture and storage.

To study environment-consumption trade-offs, we invert the abatement production function, just as we did with the grade and grain production functions in Unit 3. Suppose that, after a given level of government expenditure on other policies and also a given level of investment, the maximum amount that people could consume in the economy, that is, if no abatement is implemented, is $500bn of goods and services. Then the feasible choices are the shaded portion of Figure 18.11.

In Figure 18.11, the vertical axis still measures the quality of the environment, but the horizontal axis now measures the goods available for consumption after abatement costs. So abatement expenditures are measured from right to left. We assume that neither the economy nor the population is growing, so that consumption per person will be proportional to the total amount of consumption.



If no abatement policies are adopted

If abatement costs are zero, the nation can have €500bn of consumption.

€50bn of abatement costs

The nation is at point X after spending this amount.

Figure 18.11 *The trade-off between consumption and environmental quality.*

The abatement choice problem now looks familiar. The policymaker wishes to select from among the alternatives on the feasible frontier. Recall from the earlier units that the slope of the feasible frontier, also known as the *marginal rate of transformation (MRT)*, is how much of the quantity on the vertical axis that results if one gives up one unit of the quantity on the horizontal axis. In the consumption-environment feasible frontier, this is the marginal rate of transformation of foregone consumption into environmental quality. The steeper (the greater the slope) the less the opportunity cost in foregone consumption of further environmental improvements.

$$\textit{marginal rate of transformation} = \frac{\textit{increase in environmental quality}}{\textit{decrease in consumption}}$$

$$= \frac{\textit{increase in environmental quality}}{\textit{increase in abatement cost}}$$

Environment-consumption indifference curves

Which point on the feasible set will the policymaker choose? How much consumption are we willing to trade off to get improved environmental quality? The answer can be found by studying the policymaker's *environment-consumption indifference curves* in Figure 18.12.

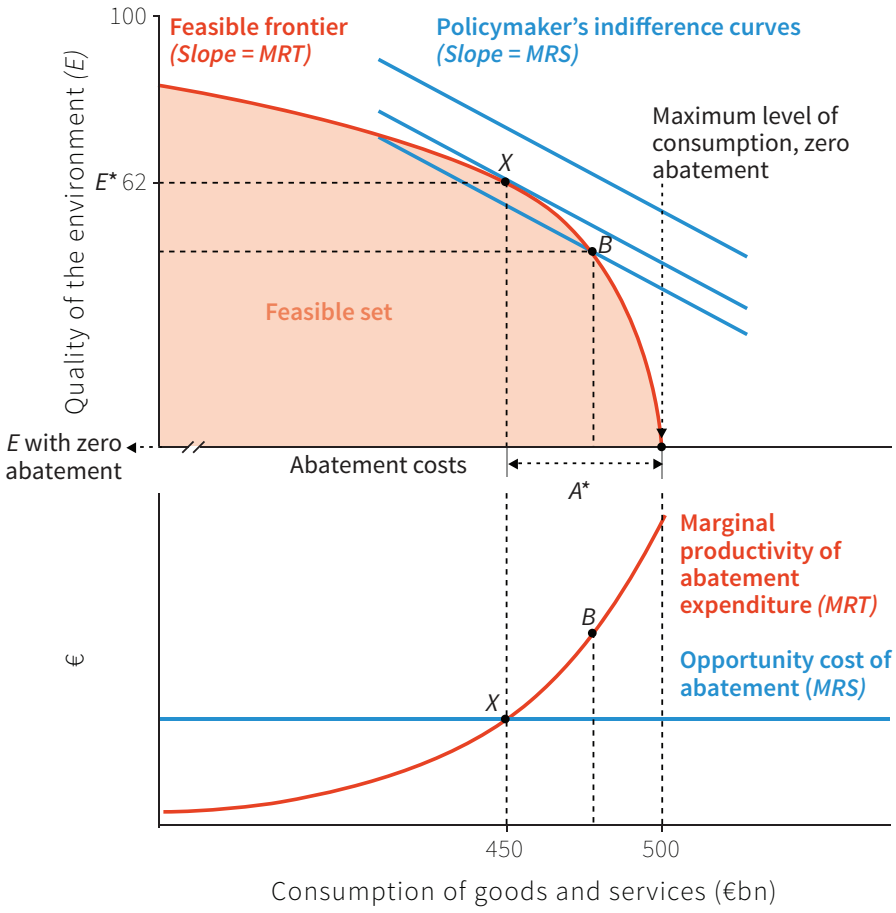


Figure 18.12 *The choice of the abatement level by an ideal policymaker.*

Although the problem looks familiar, there are two differences that we need to keep in mind:

- *Environmental quality is a public good:* It is the same for everyone (for example the effect of the abatement of CO_2 emissions).
- *The costs of abatement are spread across the population:* In our example with identical citizens, each pays $1/n$ of the total cost of abatement.

To think about what an ideal policy would be, we suppose that the policymaker takes account of the preferences of all of the citizens, counting them equally. This means that if citizens decide to value environmental quality more, then the indifference curves of the policymaker will reflect this.

We can write the slope of the indifference curve, the *marginal rate of substitution (MRS)* as:

$$\begin{aligned} \textit{marginal rate of substitution} &= \frac{\textit{increase in environmental quality}}{\textit{decrease in consumption}} \\ &= \frac{\textit{marginal utility of consumption}}{\textit{marginal utility of environmental quality}} \\ &= \frac{\textit{marginal disutility of abatement spending}}{\textit{marginal utility of environmental quality}} \end{aligned}$$

In Figure 18.12, the indifference curves are straight lines because we have assumed for simplicity that the marginal utility of consumption and the marginal utility of environmental quality are both constant; that is, they do not depend on the quantity of consumption or environmental quality. We have done this because it makes it easier to discuss the MRS if it is constant.

The policymaker's MRS will be high (a steep indifference curve) if the consumption foregone was valued highly by the citizens (a large marginal utility of consumption) and the environmental quality that is sufficient to compensate for the loss of consumption is not highly valued (marginal utility of environmental quality is low).

From this definition of the slope of the indifference curve we can see that, if abatement imposes a large cost on the citizen, the policymaker's MRS will be greater and the curve steeper. If the citizen values an improved environment more, the MRS will be less and the curve less steep. To show how we make the calculations that allow us to sketch the indifference curves in Figure 18.12, see the Einstein section.

The ideal policymaker chooses an abatement level

Our policymaker uses two principles to make a decision about the level of abatement:

- *She considers only abatement policies on the frontier of the feasible set:* This eliminates higher-cost abatement policies that are inside the shaded area.

- *She chooses the combination of environmental quality and consumption that puts her on the highest possible indifference curve.*

To satisfy both conditions, she finds the point on the feasible frontier that equates the MRT (the slope of the feasible frontier) and the MRS (the slope of her highest possible indifference curve).

We can see from Figure 18.12 that point *X* (allocating $50bn to abatement) is the level of environmental protection that the policy maker will wish to implement. This policy implies giving up €50bn of consumption to achieve environmental quality of 62 (on this index).

The second panel of Figure 18.12 shows the same information as the top panel, but now expressed in terms of the slopes of the feasible frontier and the indifference curves.

- *The marginal productivity of abatement expenditures:* This is the slope of the feasible frontier (MRT)—the marginal rate of transformation of abatement costs into improved environment. Remember: this is how much environmental improvement can be accomplished by devoting one unit of output not to consumption, but instead to abatement.
- *The opportunity cost of abatement expenditures:* This is the slope of the policymaker's indifference curve (MRS)—the marginal rate of substitution of consumption for environmental quality. Remember: this is the value the policymaker places on the consumption of goods that the citizens will have to give up if abatement policies are adopted, relative to their enjoyment of environmental quality.

In the bottom panel we can see that the marginal productivity of abatement is equal to the opportunity cost of abatement at point X. We can also see that with a lower level of abatement, indicated by point *B*, there are welfare losses due to insufficient abatement. At *B*, the marginal productivity of abatement is greater than the opportunity cost of abatement: this indicates that resources should be switched into abatement until the MRT is equal to the MRS at point *X*.

What would produce a different choice of abatement level?

- *Different values:* If the citizens cared less about the environment than the curves shown in Figure 18.12 indicate, then the indifference curves would be steeper at each level of abatement. From the lower panel, we can see that this would shift the opportunity cost of greater abatement up and imply that the policymaker would optimally choose a policy with a lower level of abatement.
- *Different costs of abatement:* If abatement became cheaper than shown in Figure 18.12, then the feasible set would be steeper at each level of abatement. From the lower panel, we can see that this would shift the marginal productivity of abatement curve up and imply that the policymaker would optimally choose a policy with a higher level of abatement.

DISCUSS 18.3: OPTIMISTIC AND PESSIMISTIC POLICIES

In Figure 18.12 we described how a policymaker representing a uniform group of identical citizens chooses the optimal amount of abatement.

1. Draw the indifference curves of the policymaker if she were to represent two different groups of citizens (again, we assume that all citizens in each group are identical). In the first group, citizens care a lot about environmental quality, and in the other group the citizens care more about consumption of goods and services. Indicate which level of abatement costs the policymaker would advocate in each case, and explain why they might disagree.

In reality, there is uncertainty about the effectiveness of abatement expenditure and hence how costly abatement of environmental damage will be.

2. On a new diagram, draw the feasible consumption frontier based on an optimistic assessment of the costs of abatement.
3. Now draw the feasible consumption frontier based on a pessimistic assessment of the costs of abatement on the same diagram.
4. By adding the policymaker's indifference curves to your diagram in each case (assuming all citizens are identical), show how actual environmental quality chosen by the policymaker will differ, depending on whether costs of abatement are assessed optimistically or pessimistically.

18.4 CONFLICTS OF INTEREST: WHO BEARS THE COST OF PROTECTING THE ENVIRONMENT?

In the previous section, we greatly simplified the problem of deciding how much abatement to do by assuming that all citizens were identical. We also invented an ideal policymaker, who even-handedly added up the benefits and costs accruing to all citizens in order to determine her preferences and the indifference curves that represented them.

Once we introduce differences among people, there are necessarily winners and losers when a society implements costly measures, or chooses to do nothing, to protect the environment.

We study two reasons for conflicts of interest:

- *Abatement costs are not equally shared among a population:* Raising taxes on automobile fuel to reduce emissions due to driving affects rural people more than urban residents, who can use public transportation. Limitations on carbon emissions by firms to protect the environment for future generations will raise costs to consumers today, and reduce the profits of the affected companies.
- *Abatement benefits are not equally shared among a population:* Environmental quality is not entirely a public good, as we assumed. We are all affected by climate change, though not to the same extent. Other environmental threats, such as living close to a factory producing toxic emissions, are local, and people with superior resources can entirely avoid localised threats.

This means there will be conflicts of interest. In this section we will continue to assume that the benefits of abatement are equally shared, but costs are not, to investigate who pays for abatement expenditures.

In our model there are two groups of people:

- *"Businesses"*: These people own and receive profits from firms whose emissions contribute to climate instability and warming.
- *"Citizens"*: People in this group make their living in other ways.

Imagine now that the businesses and the citizens are both trying to influence environmental policy. To see how they would want the policymaker to adopt different policies, let's consider what the policymaker's indifference curves would look like if she were to represent only the businesses (labelled "Businesses' indifference curves" in Figure 18.13) or only the citizens (labelled "Citizens' indifference curves").

We assume that a larger share of the costs of abatement when the policy is implemented will be paid by the businesses currently profiting from the external effects that they freely impose on all members of the society in the absence of abatement policies. This could occur because of the implementation of what is called the *polluter pays principle*.

If the polluter pays, the opportunity cost of abatement in terms of reduced consumption is higher for the business because it pays a higher share of abatement costs. To see how this affects the indifference curves, recall that:

$$\textit{marginal rate of substitution} = \frac{\textit{marginal disutility of abatement spending}}{\textit{marginal utility of environmental quality}}$$

Having to pay the costs of abatement makes the disutility of abatement spending greater for the businesses than it is for the citizens. This means that at any combination of environmental quality and consumption, the MRS is larger for the business than it is for the citizen.

$$MRS^{business} > MRS^{citizen}$$

So the indifference curve is steeper for business, as we can see in Figure 18.13. The result is that the level of abatement chosen by the citizen (at point Z) is greater than that chosen by the business (at point Y). To see how to draw indifference curves when the costs of abatement are not equally shared, see this unit's Einstein section.

Thinking of the lower panel in Figure 18.12, the opportunity cost curve of the business would be higher, leading to the selection of a point like *B* with lower abatement.

The policy adopted when society is composed of groups with two differing levels of preferred abatement will depend on which group has the greater power to influence the policymaking process. The ideal policymaker in the previous section would simply have added up the preferences of all of the owners and all of the citizens.

But this is not how the conflicting interests of citizens, business and others come to bear on public policy. Court cases, competition for political office and bargaining among the affected parties are all involved, as the next two examples show.

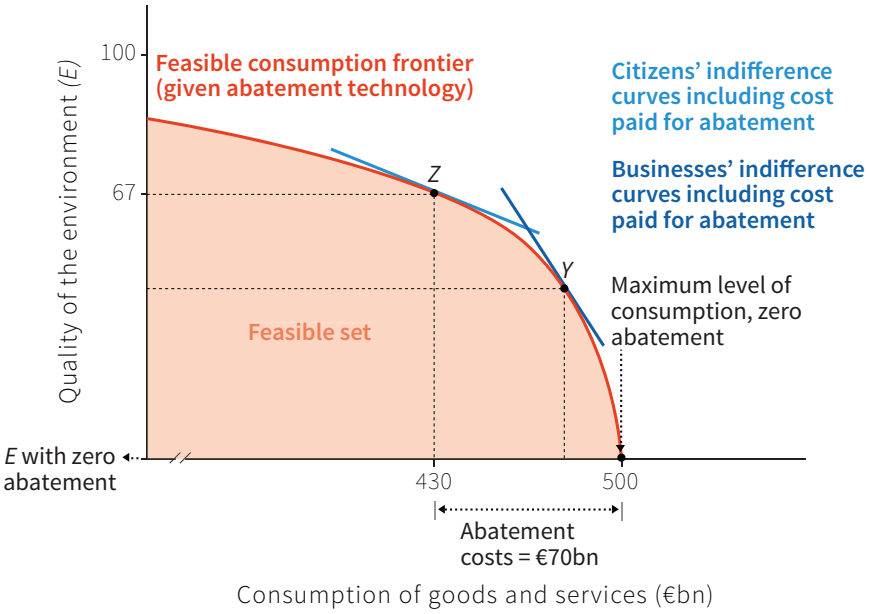


Figure 18.13 *The trade-off between consumption and environmental quality: Conflicts of interest when a policymaker represents businesses and citizens.*

18.5 CONFLICTS OF INTEREST: WHO BEARS THE COST OF A DEGRADED ENVIRONMENT?

The other conflicts of interest arise because environmental quality is never truly a public good. Some benefit or suffer more than others, depending on their location and income.

Here are two examples of how costs and benefits are not equally shared. In 2008 and 2009, two oil spills in the Niger River delta, resulting from the activities of the Royal Dutch Shell Company's extraction of oil, destroyed fisheries. Lawyers for the Ogoni people who suffered these external effects brought a lawsuit against the company in British courts, because the company is headquartered in the UK. In 2015, Shell settled out of court and paid £3,525 per person, of which £2,200 was paid to each individual, and the rest to support community public goods. This is more than the Ogoni people would earn in a year. Lawyers representing the community helped to set up bank accounts for the 15,600 beneficiaries.

The transfers may have compensated the Ogoni for the loss of their environment. For Royal Dutch Shell, the settlement at least partially internalises the external effects of their policies, and might lead the company's owners (and others extracting oil in the delta) to consider a change in policy.

In 1974 a giant lead, silver and zinc smelter owned by the Bunker Hill Company was the only major employer in the town of Kellogg, in the American state of Idaho, employing 2,300 people. Many children in the town developed flu-like symptoms. Doctors discovered that they were the result of high lead levels in their blood—high enough to impair cognitive and social development of the child.

Three of the children of Bill Yoss, a welder at the smelter, had been found to have dangerously high levels of lead poisoning. "I don't know where we'll end up," he told a reporter, "We may pull out of the state."

The company refused to release its own tests of the smelter's lead emission levels. Unless the state's emissions regulations were relaxed, it said, the smelter would shut down.

The smelter closed in 1981. Former employees looked for work elsewhere. The value of the homes and businesses in the town fell to a third of its earlier level. The local schools—supported by property taxes—did not have the funding to cope with those who remained.

We model this problem by considering a hypothetical town, Brownsville, with a single business that employs the entire labour force but whose toxic emissions are a threat to the health of the citizens. The firm can vary the level of emissions that it imposes on the town, but at a cost of capture and storage that means lost profits. The single owner of the firm (who bears the costs of reducing the level of emissions) lives far enough away that the level of emissions he selects does not affect the quality of his environment. Therefore citizens and the business will have a conflict of interest over environmental quality (the level of emissions) in the town. They also have a conflict over the wages paid.

The citizens of the town have some bargaining power because each is free to leave Brownsville and seek employment elsewhere. So the business must offer them a package of environmental quality and a wage that is at least as desirable as their reservation option, which is what they might expect were they to take their chances elsewhere. We call this limit on what the business must offer the citizens the "leave-town condition".

The business owner has bargaining power, too, because the wage and environment package that he offers must result in profits high enough that the firm does not simply shut down or relocate (we call this the "firm's shut-down condition"). The citizens cannot demand more than this wage, or they would be unemployed (there are no other firms in Brownsville). Thus the firm's reservation option places limits on the bargain that the citizens can strike with the firm.

We represent the relationship between the business and the citizens in Figure 18.14. The wage paid to the employees of the plant is on the horizontal axis. The level of environmental quality experienced by the citizens is on the vertical axis.

- *Citizens are identical and so experience the same environmental quality:* For the citizens, environmental quality is a pure public good.
- *The owner is not affected by the level of pollution:* For him the environmental external effects resulting from his decision about emissions are borne by others. Pollution for him is a private good, and he does not consume any of it.

You will probably have noticed that this figure is very similar to Figure 5.9a, in which Angela the farmer and Bruno the landowner were bargaining over the amount of grain that would be transferred to Bruno.

Here the conflict is about the amount of emissions that the townspeople will suffer. The company's profits depend on the emissions, and profits are greater if it can dispose of more toxic materials freely.

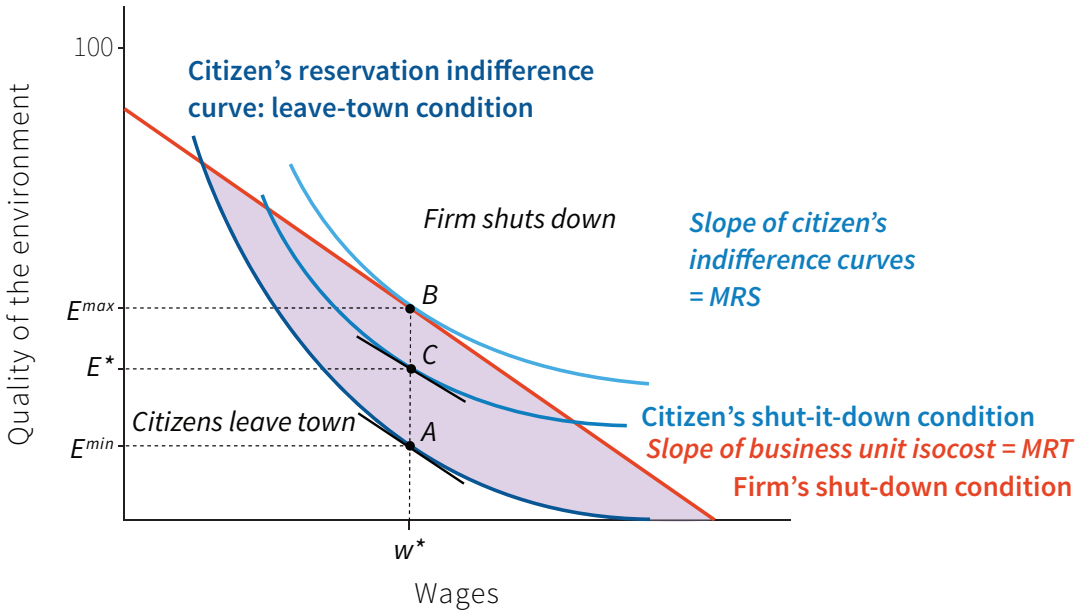


Figure 18.14 *Conflicts of interest: Whom does pollution hurt?*

The citizen's reservation indifference curve gives all the combinations of wages and environmental quality that would be barely sufficient to induce the citizen to stay in Brownsville (we call this the representative citizen's "leave-town condition"). Its position depends on what the citizen would expect to get in some other location. If she could find a high-paying job in a non-toxic community it would be higher and to the right of what is shown, for example. Its slope—the marginal rate of substitution—is the citizen's marginal utility of higher wages, divided by the marginal utility of environmental quality.

$$\text{citizen's } MRS = \frac{\text{increase in environmental quality}}{\text{decrease in wage}}$$

$$= \frac{\text{marginal utility of wage}}{\text{marginal utility of environmental quality}}$$

We assume that the citizen's marginal valuation of improvements in the environment is constant but (in contrast to the previous model) a citizen has diminishing marginal utility of receiving higher wages. At high wages (and very poor environment) on the far right of the reservation indifference curve, the MRS is small (the line is almost flat) because citizens would not care much about wages (they are already getting paid a high wage) but they are very concerned about the poor environment. At low wages the curve is steep, because they place a high value on wage increases.

The firm's shut-down condition shows the combinations of wages and environmental quality offered by the firm that would barely keep the firm in Brownsville. All of the points on this line have the same cost of producing a unit of output and, as a result, the same profit rate. It is like the isocost curve in Unit 2, and the isoprofit curve in Unit 6.

$$\text{business's MRS} = \frac{\text{decrease in environmental quality}}{\text{increase in wage}}$$

$$= \frac{\text{marginal cost of a higher wage}}{\text{marginal cost of environmental quality}}$$

The cost of raising the wage by \$1 is \$1. The cost incurred by the owner if he reduces emissions (per unit of improved environment) is p. So the $MRS = 1/p$. If the line is steep this is because p is small—avoiding emissions and thereby allowing a healthier environment is cheap.

The firm faces a trade-off: if it is at point B in the figure, it pays wages and produces emissions at a level that makes it barely profitable enough to stay in business. Therefore, if it offers a higher-quality environment to the citizens, it can only do this by offering a lower wage to them too. The opportunity cost of one unit of a better environment is p in reduced wages.

The portions of the figure above the firm's shut-down condition and below the citizen's leave-town condition are infeasible. But any combination of wages and environmental quality in the shaded portion of the figure is a feasible outcome of this conflict.

We cannot say which feasible outcome will occur, though, unless we know more about the bargaining power of the citizens and the company.

The firm has all the bargaining power

If the company could simply announce a take-it-or-leave-it ultimatum then it would find the wage and environmental quality package that minimised its costs while not violating the leave-town condition. To do this it would find the point on the citizen's reservation indifference curve at which the vertical distance between the firm's shut-down condition and the citizen's leave-town condition was the greatest. This will occur when:

$$\text{business's MRS} = \frac{1}{p}$$

$$= \frac{\text{marginal utility of wage}}{\text{marginal utility of environmental quality}}$$

$$= \text{citizen's MRS}$$

This is point A in Figure 18.14. The firm will offer a wage w^* and environmental quality E^{min}. The firm's costs will then be well below the shut-down level of costs in this case for they will be freely emitting sufficient toxic materials to reduce the citizen's environmental quality from E^{max}, the least emissions (and highest quality) consistent with the firm staying in business, to E^{min}. This difference (E^{max} - E^{min}) shows up as cost reductions, and hence profits, in the company's accounts. It also shows up as exposure to health hazards in the medical records of the people who live in the town.

Citizens have all the bargaining power

What if the bargaining power had been reversed? Suppose the citizens could impose a legally enforceable level of environmental quality in the town. What level would they impose? To determine the answer we use the citizen's indifference curves and ask: What is the highest indifference curve the citizens could be on without losing their jobs, that is, while avoiding the firm shutting down? We can see from the figure that they would impose E^{max}.

Dividing the mutual gains

The difference between E^{max} and E^{min} is a measure of the extent of mutual gains the townspeople and the business may enjoy. Any outcome between *A* and *B* on the figure is preferable to the next best alternative for the business (shut down) and the citizens (leave town). You can think this as the cake that the citizens and the business owner will divide. How these mutual gains are divided up between the two depends, as we have seen in Units 4 and 5, on the bargaining power between the two.

A point such as *C* in Figure 18.14 might be possible if the citizens, acting jointly through their town government, imposed a legal minimal level of environmental quality for the business to continue to operate. Acting together, the citizens would have more bargaining power than they get if they used the threat to leave town as individuals: they could require that the business meet at least the citizens' "shut-it-down condition" shown in Figure 18.14.

Thus bargaining power in this case would be affected by not only by the two parties' reservation options but also by:

- *Enforcement capacity:* The town government may not have enforcement capacities to impose an emissions limit on the company.
- *Verifiable information:* The citizens may not have sufficient information about the levels and dangers of emissions to win a case in court. In this case, an agreed-upon emissions level would not be complied with by the company or enforced by the town.
- *Citizen consensus:* If the town's citizens were not in agreement about the dangers of the emissions, the elected officials of the town who legislate an emissions limit might not be re-elected.
- *Lobbying:* The business may be able to convince the citizens that their health concerns were misplaced, or had little to do with the company's emissions.
- *Legal recourse:* The company may be legally entitled to emit any level of emissions that it finds profitable (perhaps subject to having purchased permits allowing it to do this).

So far we have focused on the question of how much abatement there should be. We have seen that there are trade-offs in selecting the preferred level of environmental abatement even if there are no conflicts of interest among the affected parties, and there are differences in the level of abatement that different parties will favour.

Now we consider a second question: how should this be accomplished?

18.6 THE ECONOMIC LOGIC OF ENVIRONMENTAL POLICIES

Remember that we want to achieve the desired amount of effective abatement at minimum cost. In Figure 18.12, the amount of effective abatement achieved at the chosen point, X, is the vertical distance E^*. The cost of abatement in terms of foregone consumption is the horizontal distance marked as A^*. There are two types of policies available:

- *Price-based policies:* These achieve E^* by using prices to change the signal about how resources should be allocated.
- *Quantity-based policies:* These achieve E^* by using bans and regulations.

To study how these policies work we introduce a new model to be used by the policymaker, who is now deciding how to implement abatement for an entire country.

PRICE- AND QUANTITY-BASED POLICIES

Environmental policies can be split into two types:

- *Price-based policies* use taxes and subsidies to influence our choices
- *Quantity-based policies* use bans, caps and regulations instead

To clarify her options, the policymaker considers a single typical citizen who values both environmental quality and his own consumption. The citizen engages in some activities that pollute the environment, producing a public "bad" of the type we studied in Unit 4 and section 18.3. The additional pollution that he contributes harms everyone to the same extent. This means that abatement (reducing the public "bad") produces a public good.

In deciding how to act, the citizen does not consider the benefits that abating his own pollution would confer on others. He considers only that he is harmed by his own polluting activities, and so he would benefit if he polluted less.

He also considers the private costs that he will incur in abating his pollution. Recall that the private cost is the cost to the private decision-makers, whether they are households deciding how to heat their homes, or business owners deciding how to dispose of pollution.

Suppose that there is a level of environmental quality that the members of a society would prefer, and which the policymaker wishes to implement (Figure 18.12). To implement a total desired abatement of E for the society as a whole, the policymaker must find a way to get the typical citizen to implement his share of this total, which we will call e.

To analyse environmental policies using this model, we now draw a new diagram with effective abatement by the citizen, e, on the horizontal axis and marginal costs and benefits of abatement on the vertical axis. They are measured in dollars (or some other monetary unit—we consider how we measure benefits and costs in units of money in the next section). While we consider only a typical citizen, we assume that all other citizens would respond in the same way.

Marginal private costs and benefits of abatement

The marginal private cost of abatement curve in Figure 18.15a gives, for every level of effective abatement on the horizontal axis, the addition to the citizen's total private cost of abatement of adding one unit of environmental quality through effective abatement. It slopes upward because the cost of additional abatement is high when abatement is already high. This reflects what we have seen in the feasible frontier where, as the level of abatement increases, the marginal cost of achieving a unit of abatement rises.

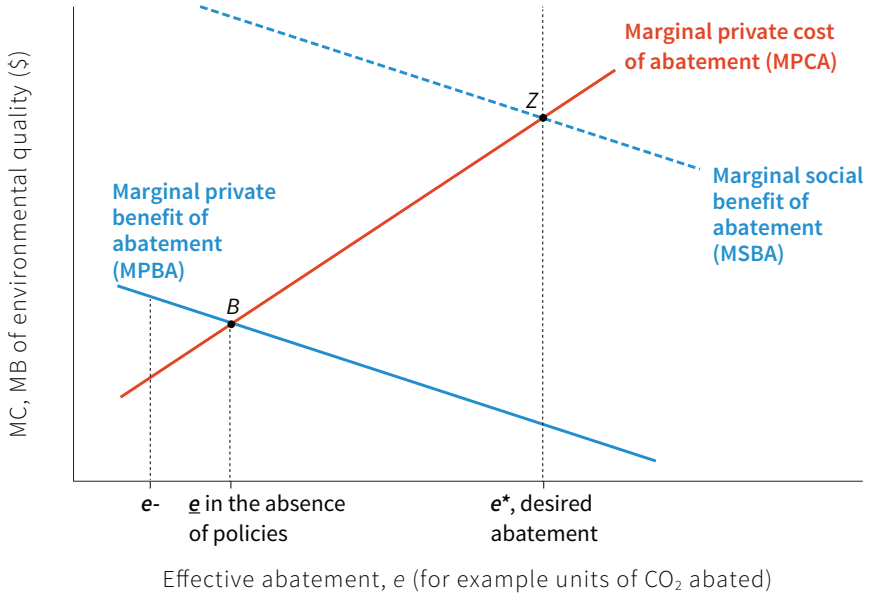


Figure 18.15a *The economic logic of environmental policy.*

The marginal private benefit of abatement curve is based on information from the indifference curves of the citizen. These differ from the indifference curves the policymaker considered in section 18.3 because they concern only the environmental benefits and costs that he experiences, not those of everyone else. He values his private benefit, namely, the contribution that his abatement will make to the

environment that he experiences. But this private benefit does not include the equivalent benefit that would be enjoyed by all other citizens. He does not take their enjoyment of a better environment into account, which is why the marginal social benefits of his abatement exceed the marginal private benefits of abatement.

The private marginal benefit of abatement curve slopes downward because the value of further environmental quality (compared to how much people value other objectives) declines as the quality of the environment improves.

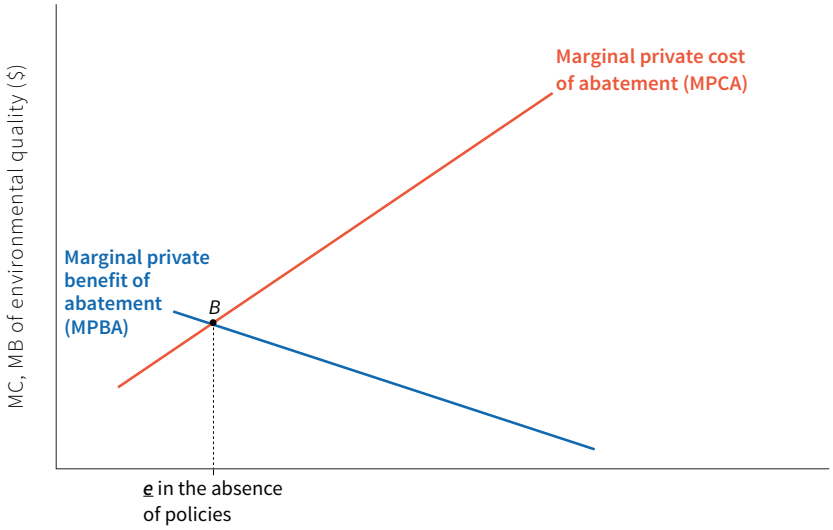
Abatement with and without environmental policies

To understand how much abatement the citizen will do in the absence of environmental policies, imagine that he were to abate at the level given by e- in Figure 18.15a, and he considers altering his abatement level. Should he abate more? Yes, we can see that the private marginal benefit of abatement exceeds the marginal cost, so he will abate more. Reasoning in this way, his private incentives lead him to abate up to the level at point B, which is well below the level that the policymaker would like to implement.

Under what conditions would he choose to implement e^*, the target amount? Just as a thought experiment, imagine that the citizen was an extraordinary altruist and valued the benefits that his abatement would confer on each of the other citizens exactly as he values his own benefits. This is shown in the figure by the marginal social benefits curve, labelled MSBA.

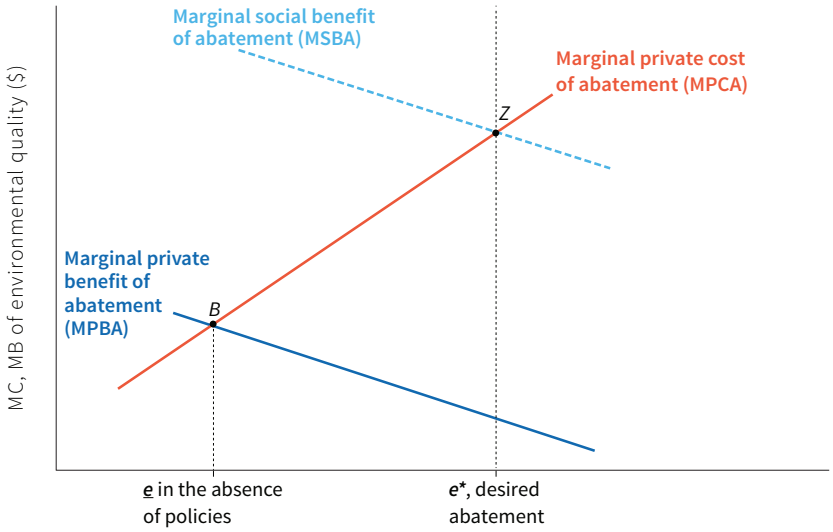
The assumption of complete altruism is unrealistic, but it allows us to see that if he were to fully internalise the benefits of his abatement actions to others (just as the ideal planner did in previous sections), the desired level of abatement would be implemented privately (that is, by his own incentives at point Z). There would be no need for the policymaker to intervene.

As we know from Unit 4, many people care about the effects of their actions on others, so we might expect the typical citizen to consider at least some of the external effect of his abatement. The policymaker would also consider using persuasion and education to make people aware of the environmental effects of their actions on others. These policies might shift the marginal private benefits curve upward, as shown by the curve labelled “effects of education, persuasion on MPBA” in Figure 18.15b.



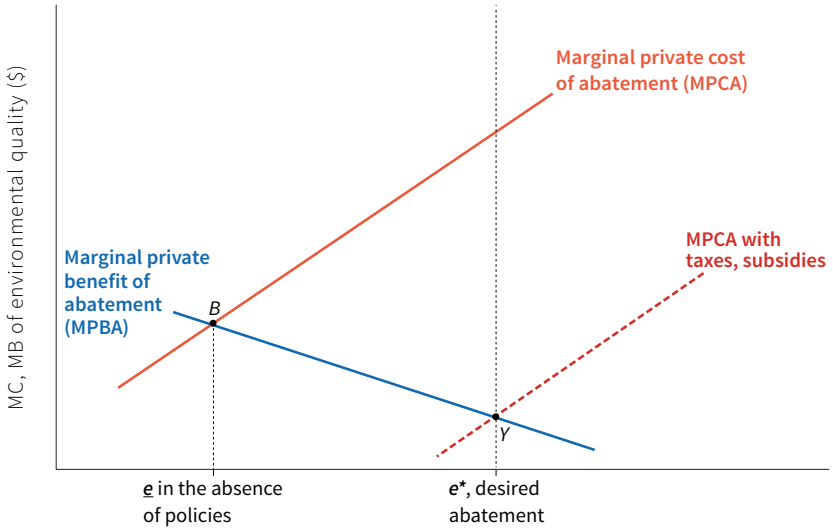
The outcome without intervention

We begin with the intersection of the private marginal benefit and marginal cost curves: this shows the outcome in the absence of government intervention (point *B*).

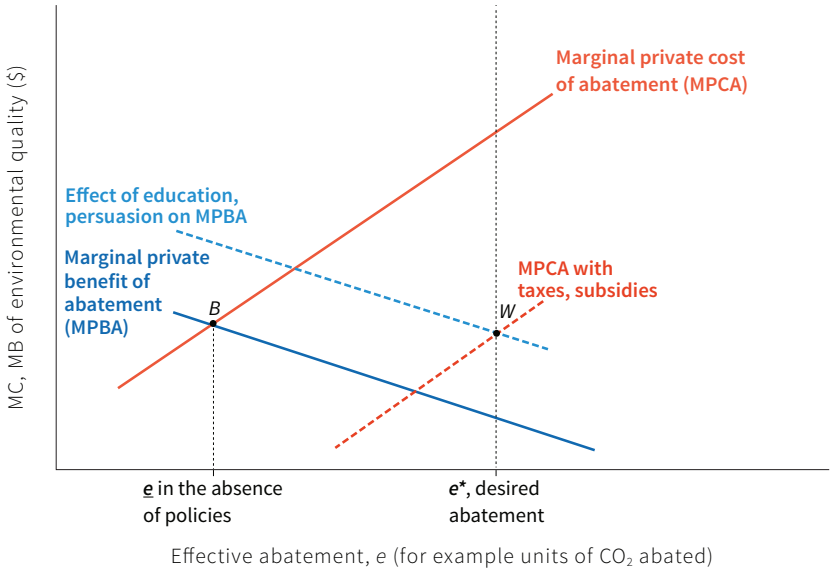


The chosen level of abatement

This could be achieved by private action if the citizen internalises the benefits to everyone of his own abatement, so the MSBA intersects the MPCA at point *Z*...



... Or by taxes and subsidies that shift the MPCA down (point *Y*)...



... Or by a combination of education and persuasion on the one hand, and taxes and subsidies on the other (for example, point *W*).

Figure 18.15b *The economic logic of environmental policy.*

Other policies can reduce the net private costs of abatement, shifting the MPCA curve downward. By *net costs* we mean:

- *The cost of the abatement itself* (such as the cost of installing and using solar panels).
- *... Subtracting the cost of whatever energy source she is now using* (for example, oil).
- *... Also subtracting any subsidy for adopting a renewable energy source that she may receive.*

Sticking with the solar panel example, policies that can reduce the net costs and shift the MPCA curve downward include:

- *Subsidies for R&D into, and production of, solar panels:* These lower the cost of the abatement technology.
- *A tax on the use of fossil fuels:* This raises the cost of the environmentally damaging technology.
- *A subsidy offered to users of solar power:* This offsets some of the private cost of using the abatement technology.

Cap and trade: Creating a market for emissions

A policy called *cap and trade* combines a quantity-based limit on emissions with the price-based approach of placing a cost on damaging production or consumption decisions.

CAP AND TRADE

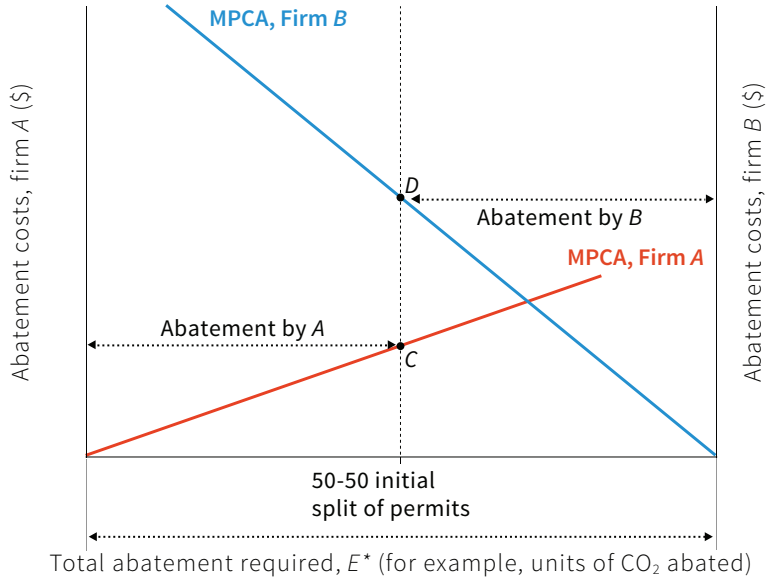
A policy through which a limited number of permits to pollute are issued, and can be bought and sold on a market. Cap and trade combines:

- A quantity-based limit on emissions
- A price-based approach that places a cost on environmentally damaging decisions

Environmental external effects arise because of missing markets. So why not create a market in which firms have to pay to emit CO_2 by buying a permit? Their incentive to abate will be increased. It's as if they were paying a tax on emissions.

To show how this works, we first find the Pareto-efficient level of emissions (or, equivalently, the total level of abatement required, E^*) using the MCA/MBA analysis in Figure 18.15b. This is shown by the length of the horizontal axis in Figure 18.16.

Work through the slideline in Figure 18.16 to see what happens if the number of permits is initially divided equally between two firms with different costs of abatement.



The MCA of firm *A*

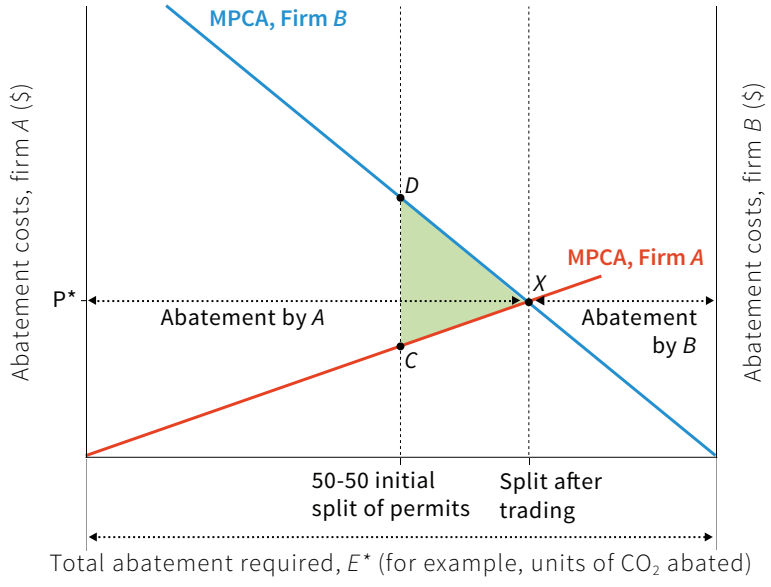
This is measured in the usual way from the left-hand axis: it rises as its cost of abatement increases.

The MCA of firm *B*

This is measured from the right-hand axis: it rises from the right origin as *B* engages in more abatement.

Permits split 50-50

Let's see what happens if the permits to pollute are initially split 50-50 between the two firms. Firm *B* has a higher MCA; this creates gains from the trade in permits. Firm *B* wishes to pollute more since its cost of abatement is so high.



Firm *B* will buy permits from *A*

A will receive the revenue from selling permits to *B* and abatement will take place at *X*, where the MCA is the same in each firm.

The gains from trade

The shaded triangle shows the gains from trade created by the market for permits. *P** is the permit price and is equal to the marginal cost of abatement in the economy.

Figure 18.16 *Hybrid policy: Tradable permits to pollute.*

The trading of permits achieves the Pareto-efficient level of abatement at least cost of resources to the economy. P* is the permit price and is equal to the marginal cost of abatement in the economy.

For cap and trade to operate successfully:

- *The government or governments set the total level of abatement required*: This is called the cap.
- *The government creates permits:* The number of permits issued allows total emissions to equal the size of the cap.
- *The government allocates permits:* They can be given to the firms operating in industries emitting the pollutant, or they can be auctioned.
- *The permits are traded:* The market-clearing permit price, P*, does not depend on how the initial permits are distributed. Trading will take place to eliminate the gains from trade.

Allocating the permits by an auction raises revenue for the government. Another benefit of an auction of permits is that the revenue can be used to reduce taxes that create distortions in the allocation of resources, such as business taxes that are based on the number of workers that firms hire. These taxes discourage firms from hiring.

Cap and trade: Examples of emissions trading schemes

One of the earliest successful emissions trading schemes was the sulphur dioxide (SO_2) cap and trade scheme in the US, implemented in the 1990s and intended to reduce acid rain. The allowances were free: the most polluting power plants received the most permits. By 2007, annual SO_2 emissions had declined by 43% from 1990 levels, despite electricity generation from coal-fired power plants increasing more than 26% during the same period.

The European Union *Emissions Trading Scheme* (EU ETS), launched in 2005, is the largest CO_2 cap and trade scheme in the world, and now covers 12,000 polluting installations across the EU. National governments auction 57% of permits in the

EU ETS, and the overall emission cap is tightened every year. Some of the auction proceeds are used to fund low-carbon energy innovation. Similar carbon trading schemes exist in other countries and regions.

The EU ETS has been less successful than the US SO_2 scheme. Some analysts think this is due largely to the fact that the permitted level of emissions was too high (too large a cap). After the financial crisis in Europe, lower aggregate demand caused the demand for electric power to shrink. Firms did not want to produce levels of output that would generate carbon above the cap, so the price of permits fell dramatically. This allowed firms to pollute without regulation and at low cost, as shown in Figure 18.17.

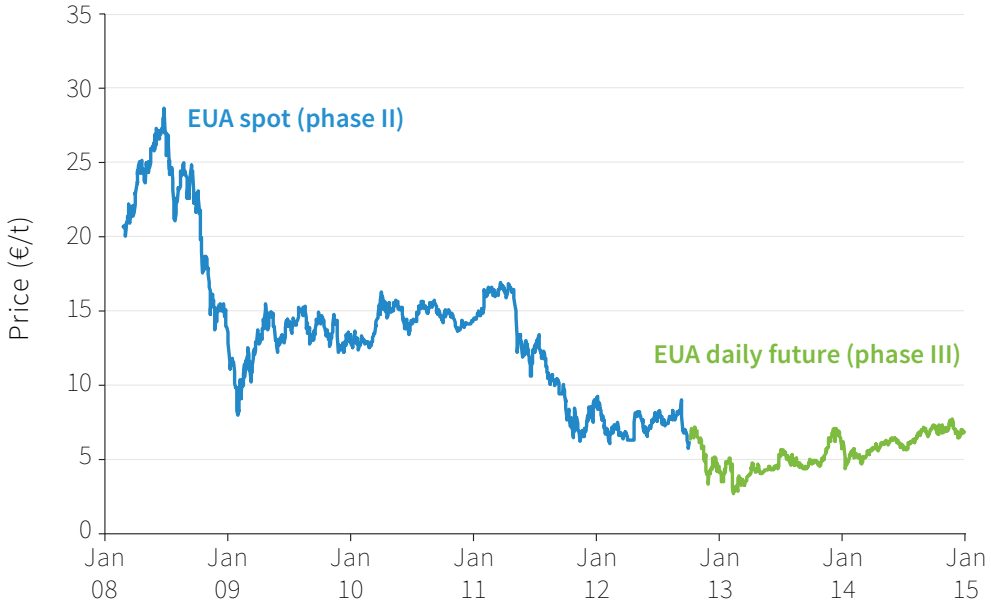


Figure 18.17 *Permit prices in the European Union Emissions Trading Scheme (EU ETS).*

Source: *Data provided by SendeCO2 based on prices from Bloomberg Business.*

Although in the short run emissions are below E^*, the reason is poor performance of the aggregate economy. This highlights a drawback of cap and trade. The price signal is not necessarily a reliable guide for future abatement investment decisions. In Germany, for example, this led to several high-emitting coal power plants re-opening, because dirty technology was profitable again.

As long as the cap is binding, a tax on carbon emissions and a cap and trade policy obtain the same outcome: the Pareto-efficient level of abatement, E^*, by setting the right price for carbon emissions. In both cases the policymaker must decide on E^* first, before selecting the most appropriate policy. Note also that emission trading schemes do not need to leave the market entirely free. The UK, for example, uses a *carbon price floor*, which sets a minimum price for British participants in the Emissions Trading Scheme.

DISCUSS 18.4: A SUCCESSFUL TRADABLE EMISSIONS PERMIT PROGRAMME

The cap and trade sulphur dioxide permit programme in the US successfully reduced emissions. The programme costs were approximately one fiftieth of the estimated benefits.

Read this article.

1. In the view of the authors, why are cap and trade systems such powerful tools to achieve reductions in emissions?

Now read this paper by Richard Schmalensee and Robert Stavins.

2. Summarise the evolution of permit prices using Figure 2 in the article.
3. How well can the price movements in permit prices be explained by the analysis in Figure 18.16?

Look again at Hayek's explanation of prices as messages (Unit 9), the analyses of asset price bubbles (Unit 9) and housing bubbles (Unit 17).

4. Could we use similar reasoning to explain price movements in Figure 2 of the paper by Schmalensee and Stavins?

18.7 MEASURING THE COSTS AND BENEFITS OF ABATEMENT

To implement environmental policies using the marginal costs and benefits framework, we need to measure the costs and benefits of abatement.

- *Measuring costs of abatement:* As we saw in Figure 18.7, this requires that we know the range of technologies used in electricity generation, agriculture or other industries that emit CO_2, and the cost of reducing emissions in each industry. The data demands for other forms of abatement range from the extremely challenging—preserving biodiversity, protecting the oceans—to the relatively routine—ensuring drinking water to urban populations, curbing acid rain. Below, a natural experiment is used to uncover one element that is needed to measure the costs of air pollution: its effects on life expectancy.

HOW ECONOMISTS LEARN FROM FACTS

THE EFFECT OF AIR POLLUTION IN CHINA

Figure 18.18 *China: the Huai River policy boundary and locations of disease survey points (1991-2000).*

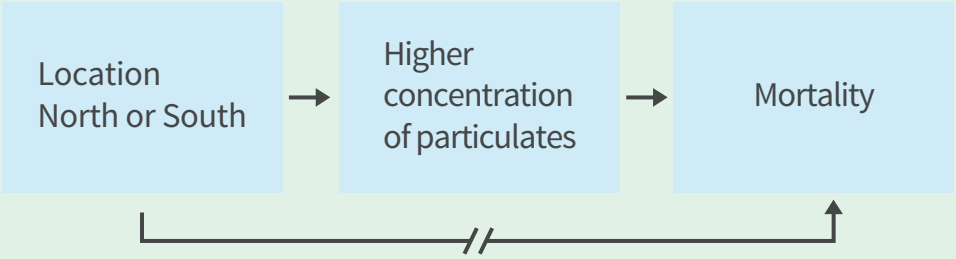
Source: *Chen, Yuyu, Avraham Ebenstein, Michael Greenstone, and Li Hongbin. 2013. 'Evidence on the Impact of Sustained Exposure to Air Pollution on Life Expectancy from China's Huai River Policy.' Proceedings of the National Academy of Sciences 110 (32): 12936–41.*

China's local air pollution is having an impact on life expectancy, but how do we estimate how much would be gained by abatement? In 2013 economists Yuyu Chen, Avraham Ebenstein, Michael Greenstone and Hongbin Li used Chinese mortality data between 1991 and 2000 to estimate that an increase in particulate concentration of 100μg/m3 leads to a decline in life expectancy of three years, mostly due to fatal heart attacks. China's particulate concentration in large cities is typically 400 μg/m3!

How did they estimate this effect? They could have collected data on particulate concentration and mortality for every city, and simply looked at whether cities with higher particulate concentration have higher mortality. But, since cities with higher particulate concentration could be systematically different from cities with lower particulate concentration—for example they could be poorer, which researchers may not be able to observe—this would not tell us the causal effect of particulate concentration on mortality.

The researchers noticed that between 1950 and 1980 the government provided free coal for winter heating to homes and offices north of the Huai river. Figure 18.18 shows the black line that formed the boundary of the policy, which follows the Huai River and the Qinling mountain range. Cheaper coal and greater heating needs meant that homes in northern China used a lot more coal, which increased concentration of harmful particulates in the air.

The Huai river policy should affect mortality only through its effect on particulate concentration: for example, other sources of air pollution are roughly similar north and south of the Huai river.



So the researchers looked at the relationship between particulate concentration as predicted only by whether a city was north or south of the Huai river (as well as its latitude and other city-specific characteristics) and mortality. This strips out all the unobservable (to researchers) things that can affect both mortality and particulate concentration, such as poverty, and allows researchers to identify the causal effect of particulate concentration on mortality. What they found was that the concentration of particulate matter was 55% higher north of the river and life expectancy was 5.5 years lower.

- *Measuring the benefits of abatement:* Placing a value on the benefits of abatement is challenging because we are dealing with missing markets for environmental quality. What is the value of preserving a wilderness, saving a threatened species, creating better air or less noise?

Economists have used creative methods to measure the benefits of abatement. We examine three here: hedonic pricing, contingent valuation, and adjusting GDP to account for the environmental external effects of production.

Contingent valuation

Among the simplest and most widely used methods of valuing the benefits of abatement is just to ask people. For example, after the 1989 Exxon Valdez oil spill in Alaska, which released 11 million gallons (42 million litres) of crude oil into beautiful Prince William Sound, the court used *contingent valuation* to assess the value of the losses (such as the value of natural beauty) caused by the spill. They did this in a survey by asking respondents how much they would be willing to pay to prevent a new spill. The study estimated the lost value in 1990 to be at least $2.8bn. Exxon eventually paid $1bn in damages in a settlement with the governments of Alaska and the United States.

Researchers used contingent valuation techniques to get a quantitative estimate of the value of elephant conservation in Sri Lanka. Farmers were killing elephants to protect crops and homes. The researchers wanted to know how much Sri Lankans

would be willing to pay to the farmers as compensation for the damages caused by the elephants, if the farmers stopped killing them. This would be a Pareto improvement: if implemented, it would make both citizens and farmers better off, or at least not worse off (not to mention the effect on the elephants).

Contingent evaluation is called a *stated preference* approach because it is survey-based and accepts the respondents' statements of their values as indicative of their true preferences. This is not the case for hedonic pricing.

Hedonic pricing

Hedonic pricing is called a *revealed preference* approach because it uses people's economic behaviour (not their statements) to reveal what their preferences are. Laboratory experiments are a similar method of studying revealed preferences, as we saw in Unit 4. But lab experiments are not very useful in valuing the environment.

An example of hedonic pricing: how much is it worth to you to not have your residence bombarded by the sound of airplanes flying overhead? Economists answer this starting with the observation that houses under aircraft flight paths are sold for less than equivalent houses in quieter locations. By comparing data on house prices, we can calculate the amount people are prepared to pay to avoid the noise pollution.

This technique was used in the UK to set the tax for landfill waste. The marginal benefits of abatement were estimated in a study that used data on more than half a million housing transactions over the period 1991-2000. By controlling for a large number of factors that can account for the variation in house prices, the researchers then tested whether any of the variation left unexplained could be accounted for by the proximity of the house to a landfill site. The researchers found that being within a quarter of a mile (400m) of a working landfill site reduced house prices by 7%. They calculated that the marginal benefit from reducing the proximity to a landfill site was £2.86 per tonne of waste (in 2003 prices).

Adjusting GDP

Environmental degradation is not explicitly measured in national accounts, yet. The World Bank estimates that natural capital comprises 36% of wealth in developing countries.

Remember that income is the most a person, or a nation, could consume without reducing its capacity to produce in the future. This was the message of the bathtub in Unit 11: income is the flow of water into the tub *minus* the amount of evaporation that is reducing the total amount of water in the tub. Income according to this definition is gross income minus depreciation.

Recall also that depreciation refers to the wearing out or using up of the capital goods used in production. But when it comes to a nation's natural capital, this is not how income is measured. The portion of a nation's capital that is used up in any year is not subtracted.

Below you will learn about how some economists are changing this by placing a monetary value on the use of natural assets.

HOW ECONOMISTS LEARN FROM FACTS

GDP MEASURED WITH ENVIRONMENTAL LOSSES

How much money is natural degradation or biodiversity loss worth? In order to take natural capital loss into account (often referred to as a *green adjustment* of national accounts) we must figure out how much it will cost (per year) to replace the lost natural capital and subtract it from the annual GDP figure. Firms routinely estimate the depreciation of their assets through wear and tear. When Indonesian government policy generated a timber boom between 1979 and 1982, Robert Repetto and his colleagues from the World Resources Institute estimated that the country sacrificed more than $2bn of potential forest revenues.

Repetto and his co-authors also showed that, considering oil depletion, soil erosion and deforestation, Indonesia's average annual economic growth rate—originally reported as 7.1% from 1971 to 1984—was in reality only 4%. The impact of natural resource destruction on GDP was calculated by assigning a monetary value to those losses (for example the cost of replacing the assets), considering the total loss as a negative investment, and subtracting it from the official figures.

A similar exercise was carried out for Sweden between 1993 and 1997 where the loss of natural assets was around 1% of GDP per year.

DISCUSS 18.5: WEALTH AND NATURAL CAPITAL

Use the World Bank data in The Changing Wealth of Nations report. Download the total wealth of nations data.

1. For 10 countries of your choice, calculate the change in natural capital between 1995 and 2000 and between 2000 and 2005 in absolute terms. Summarise and interpret your results.

Go to The World Bank data. Find and download GDP (in constant prices) for your chosen countries for 1995, 2000 and 2005.

2. Calculate the change in GDP between these periods. You may want to draw a scatter plot that compares the two sets of data. Does it look like there is a relationship in the data between the change in GDP and the change in natural capital for these countries?
3. Suggest explanations for any relationship you find.

WHEN ECONOMISTS DISAGREE

WILLINGNESS TO PAY VERSUS THE RIGHT TO A LIVEABLE ENVIRONMENT

The Constitution of the Republic South Africa enshrines the citizen's "right to an environment which is not detrimental to his or her health or wellbeing". The Supreme Court of India ruled that the "right to life" guaranteed by the Constitution of India "includes the right to enjoyment of pollution free water and air…" Similar rights are granted in at least 13 other constitutions, including Portugal, Turkey, Chile and South Korea. Use this web site to check the constitution of your country, or any other in which you are interested, to see if you can find these guarantees.

Political movements opposing the privatisation of water supply have evoked similar language: access to clean water, they argue, is a human right.

When a feature of the environment such as proximity to a landfill, noise pollution, or toxic emissions from a smelter is valued in monetary terms using the methods described above, this ignores the principle advanced by many that people have a right to an environment free of these hazards.

But in response, others ask: why should the quality of the environment that you experience be any different from the quality of the car that you drive or the food that you eat? You get what you pay for, and if you are unwilling to pay, then why should the policymaker worry about your values? If you believe this, the benefits of abatement policies can be measured by the citizens' willingness to pay for the improved environment that the abatement will allow.

The willingness to pay measure is criticised by some economists and citizens because it implies that people with hardly any money place a limited value on the environment, just as they have a limited willingness to pay for anything else. It is not that they lack the will; they lack the way. Therefore using willingness to pay as the method of estimating the benefits of abatement—for example, when either contingent valuation or hedonic pricing is used—means that policies that improve environmental hazards that mostly affect the poor, like ensuring safe drinking water in urban areas, will be valued less than policies that raise the environmental quality experienced by rich people, like pristine rivers, lakes and oceans to enjoy while boating.

If a safe environment is a right, an economist would term it a *merit good*, which you may recall from Unit 10. It is like the right to vote, or legal representation in court, or an adequate education: a good that should be available to all citizens irrespective of their wealth.

The advantage of the approach based on willingness to pay is that it makes use of information on how people value the environment. This should be relevant to how much we invest in environmental quality. Defining the environment as a right has the advantage that it does not give priority to the preferences of those with higher incomes in shaping environmental policy.

18.8 ENVIRONMENTAL DYNAMICS: FUTURE TECHNOLOGIES AND LIFESTYLES

The trade-offs given by the feasible sets and indifference curves we have used in our analysis will change as people adopt new values and lifestyles and develop new technologies, and as our impact on the environment intensifies. Our discussion of the economic logic of environmental policy made it clear that a policymaker's objectives include changing people's preferences and improving the technologies that define what is feasible today.

Prices, quantities and green innovation

Improvements in technology can enlarge the feasible set. Some improvements may make abatement more efficient, lowering the opportunity cost of an improved environment. Others may improve methods in producing other goods, reducing the environmental costs of more consumption as a result. Figure 18.19 illustrates the effect of a technological improvement in abatement, which improves the marginal rate of transformation of foregone consumption into improved environment. By increasing the marginal productivity of abatement expenditure, it makes the feasible frontier steeper. This would appear in Figure 18.15 as a shift downward in the marginal cost of abatement.

DISCUSS 18.6: AN IMPROVEMENT IN TECHNOLOGY

Redraw Figure 18.19 showing an improvement in the technology for producing consumption goods, and show the new combination of the two goods chosen by the citizen.

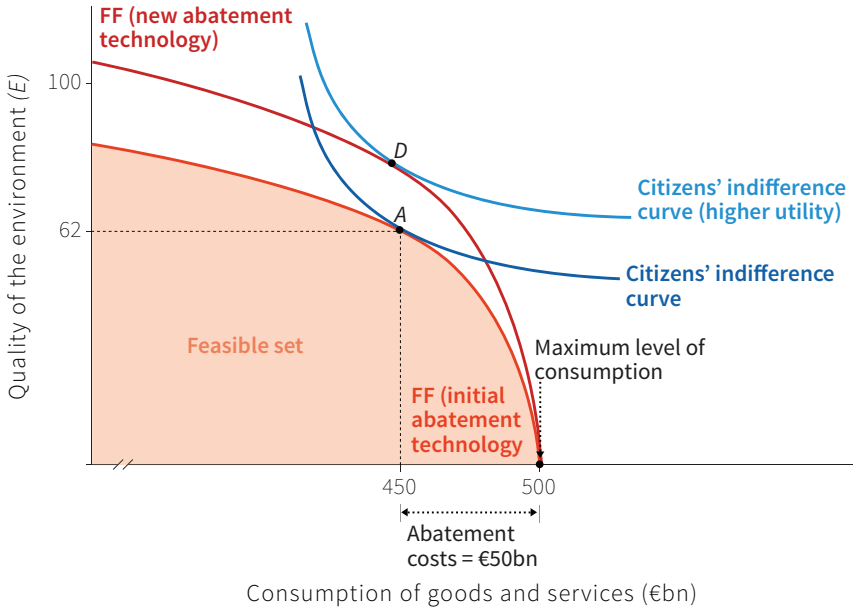


Figure 18.19 *The abatement technology changes.*

In Unit 2 you learned how the rents from innovation drive progress and the improvement of productivity. If the right incentives exist to create innovation rents, we would expect technological breakthroughs that can deliver substitutes for some resources that would be used up, or that need to stay in the ground if temperature increase is to be safely contained. One such case is the technological progress achieved in solar energy.

Increased use of solar power by firms, with subsidies to firms producing the panels and other equipment, has resulted in fast declines in the cost of generating solar power. Figures 18.20a and 18.20b show that, over the last few decades, we have seen a dramatic improvement in photovoltaic cell efficiency, which implies a reduction in the cost of producing solar electricity. Already in the United States many renewable energy technologies can compete with fossil fuel generation, in terms of the cost of new electricity generation capacity, without subsidies. (Note: we can only generate wind power when the wind blows, and solar power when the sun shines, which makes them harder to integrate into the energy system. It is likely that the electricity system of the future will need several renewable technologies side-by-side as well as plenty of energy storage.)

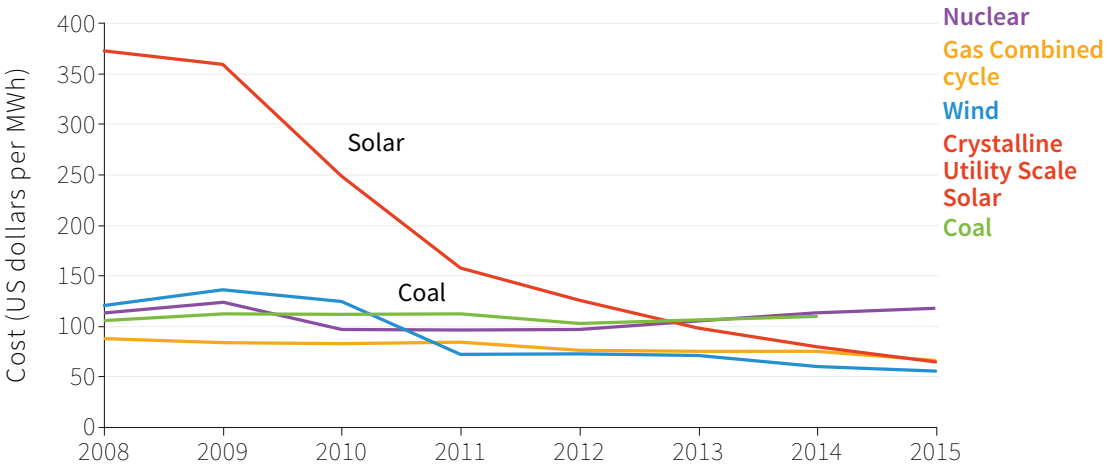


Figure 18.20a *Cost of generating electricity (new capacity) from different sources in the US (2008-2015).*

Source: *Lazard. 2015. 'Levelized Cost of Energy Analysis 9.0.' Lazard.com. November 17.*

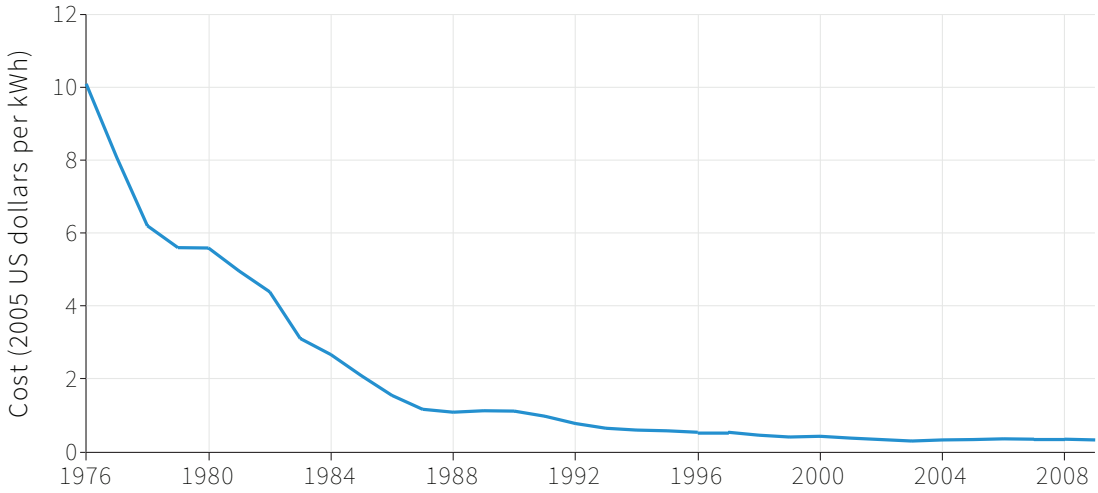


Figure 18.20b *Cost of generating electricity (new capacity) using photovoltaic cells in the US over time.*

Source: *Nemet, Gregory F. 2006. 'Beyond the Learning Curve: Factors Influencing Cost Reductions in Photovoltaics.' Energy Policy 34 (17): 3218–32; Nagy, Béla, J. Doyne Farmer, Quan M Bui, and Jessika E Trancik. 2013. 'Statistical Basis for Predicting Technological Progress.' PLoS ONE 8 (2). Public Library of Science (PLoS).*

After the oil crises of the 70s, many oil-dependent countries spent public resources on basic research in renewable energy. Now, following 20 years of neglect, support for basic energy research is back on the public agenda. As governments aim to promote positive spillovers from innovation and learning in renewable energy, which technologies should they support? Should they let costs of technologies come down further before they give them more support, or should they pick winners early? As you will see in Unit 20, technological breakthroughs can be unpredictable and mistakes can be costly, but government subsidies for basic research can accelerate the pace of technological change.

To illustrate how a tax can create innovation rents by changing relative prices and promote innovation by the private sector, we apply a model introduced in Unit 2. Imagine a textile producer called Olympiad Industries (a hypothetical business), located in a country where the supply of electricity is intermittent, and so like most firms in the country it owns a coal-fired power generator. Burning fossil fuel generates greenhouse gases but the alternative (solar power) is more expensive. While the firm has installed some solar panels, it relies primarily on coal for electricity generation.

Figure 18.21 illustrates the cost comparison. You will be familiar with the model: it is the one in Unit 2 in which we explained how relatively high wages in England made the introduction of a labour-saving innovation—the spinning jenny—profitable. The difference is that we are not considering an innovation that saves labour but instead one—solar energy—that saves environmental resources many of which (unlike labour in England in the 18th century) have no price.

$$MRS \text{ of less free time into more air travel} = \frac{\text{increase in air travel}}{\text{decrease in spare time}}$$

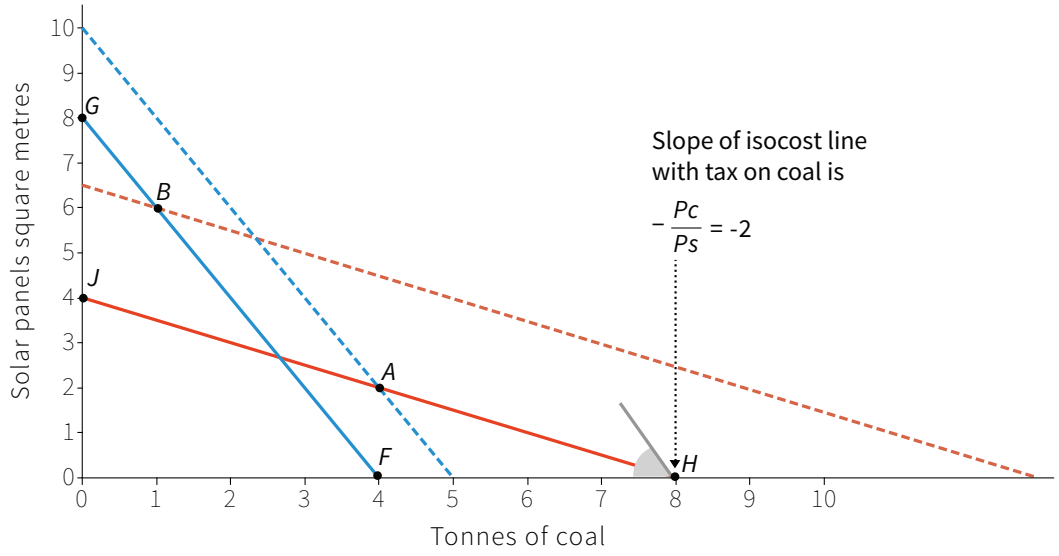


Figure 18.21 *Olympiad Industries' choice: The effect of an environmental tax on firm behaviour. Coal- versus solar-intensive local power generation technology for textile production.*

The point A represents Olympiad's current technology, at the market price of coal and solar power. They are using 4 tonnes of coal and 2m2 of solar panelling to produce power sufficient for 100m of textiles.

There is an alternative technology represented by point B using almost entirely solar power with just a bit of coal use for periods of the year when solar is unreliable. The isocost line is shown in red. It indicates all of the possible combinations of solar and coal (sufficient to produce 100m of textiles) that have the same cost. Isocost lines closer to the origin represent lower costs. The flat slope of the isocost line says

that coal is a bargain. The solar alternative (point *B*) is on an isocost line indicating a higher cost for producing the same level of output than coal, which is why the owner of Olympiad has decided to stick with coal despite his concerns about climate change.

But now suppose that the environmental policymaker has imposed a tax on electricity produced using fossil fuels. This means that for the same cost as 4 tonnes of coal, the company could now be using 8 solar panels. The new blue isocost line shows that *B*, the solar alternative, is now cheaper than *A*, the status quo coal-based technology, for producing 100m of textiles. The dashed blue line represents the isocost line after tax for which the firm has the same cost as using input combination *A*. Now you can see that the new isocost line through *B* is now inside (a lower cost) the dashed blue isocost line through *A*.

This gives the owner of Olympiad a reason to adopt solar technology. Here the tax has changed the message sent by prices. It now says that you can make a profit by using renewable sources of energy. It also says: sticking with coal may mean being undercut by your competitors, if they switch to the lower-cost technology.

Environmental policy and long-term changes in a way of life

In the long run, in addition to the role of policy in green innovation, how much we value the goods that contribute to our wellbeing can also change. Environmental and other policies can contribute to changes that reduce the negative impacts of our choices on the environment.

In Figure 3.1 you saw that production workers in the Netherlands worked much less than half as many hours in the year 2000 as they had in 1900. In 2000 they enjoyed a lot more free time and consumed less than half as many goods and services as they would have done had they continued working more than 3,000 hours a year, as they did in 1900. Were they still working long hours and consuming twice as much as they do now, their adverse impact on the environment would be larger.

Look ahead to Figure 18.25a, which shows the CO_2 emissions and GDP per capita for a wide range of countries. As a thought experiment, imagine that the Netherlands were twice as rich as it is in that graph. What would be the environmental impact in terms of CO_2 emissions? In that figure the Netherlands is slightly below the "predicted" line and so if we assume that this was also true of our hypothetical workaholic Dutch nation, we can determine the level of CO_2 emissions using the predicted line. Instead of emitting 11 tonnes of CO_2 per capita per year, they would be emitting more than 20 tonnes. This would make the Netherlands among the top polluters in the world.

The Netherlands experienced an unusually large fall in its work hours (Figure 3.1 shows that work hours in France and the US fell, but not on the Dutch scale). But even for these and other countries, had free time not expanded at the opportunity cost of less consumption, the impact on global climate change would have been worse.

A lifestyle that is rich in free time, and less rich than it could be in goods and services produced in the economy, is a "greener" lifestyle. Environmental policies can contribute to people adopting this lifestyle.

To see how, imagine that Omar is considering how much air travel to do on his holiday. Omar has enough income to fly anywhere, but he knows that burning aviation fuel is a major source of greenhouse gases. He would also like to have more free time, but realises that a shorter working week would mean he has less money for his next holiday.

We represent the trade-offs affecting his choice in Figure 18.22. On the horizontal axis we measure hours of free time per year that he would have, if he worked just long enough so that he could pay for all of the other things he spends money on (clothing, rent, food, and entertainment). On the vertical axis we indicate his kilometres of air travel during the year. The red line gives the total amount of air travel that he can afford for each of the hours of free time (the shorter work week) that he might select. So the red line is his feasible air travel-free time frontier.

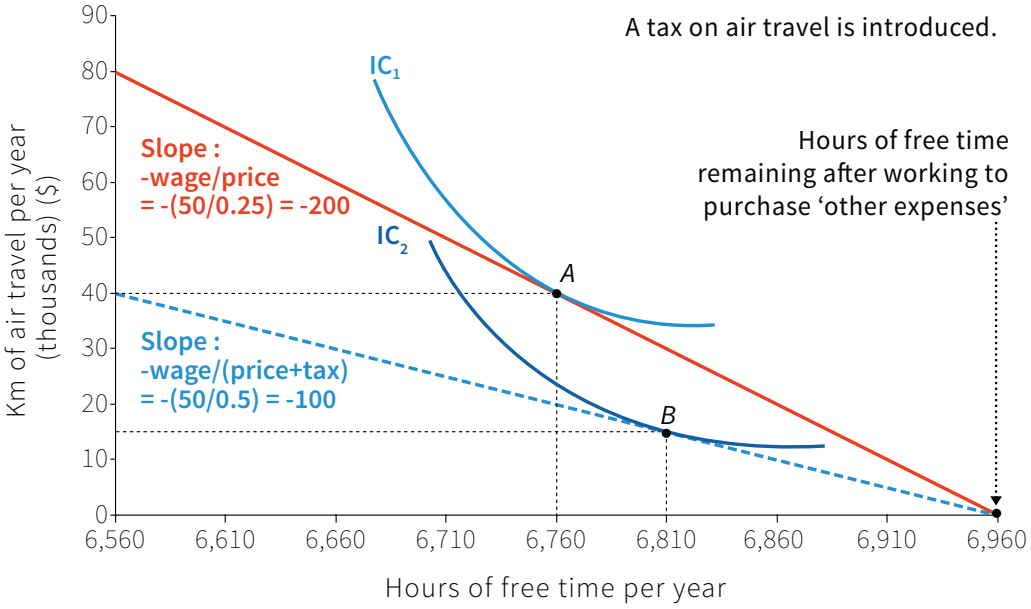


Figure 18.22 *Omar's choice: The effect of an environmental tax on consumer behaviour as air travel and free time.*

The feasible frontier is constructed as follows. Suppose Omar makes $50 an hour after taxes and that he is free to set his own hours of work. He spends $90,000 on things other than air travel and, to earn this amount, he must work 1,800 hours during the year. So, from the 8,760 hours in the year that he could give to work (as in Unit 3), he chooses to work 1800 hours. Thus he has 6,960 hours of free time if he takes no air travel at all: this is the horizontal axis intercept of the frontier. How much air travel will he choose if $1 buys 4km of air travel for the kinds of trips Omar would consider making?

DISCUSS 18.7: THE PRICE ELASTICITY OF DEMAND

A study of vehicle use and gasoline prices in California estimated that the short-run price elasticity of demand for the number of miles a car is driven is -0.22. Suppose the price of gas is now $3 per gallon and a proposed tax would raise the price to $4 per gallon.

1. What is the predicted reduction in the miles driven if the tax is implemented?

The same study found that people with higher incomes responded more to gas price changes than people with lower incomes.

2. Can you think of reasons why this may be the case?
3. Sketch two demand curves: one for high-income people and one for low-income people. Show why the tax will impose a larger cost on the low-income group.

To answer this question we have to ask: what is the MRT of foregone free time into feasible air travel? This is the slope of the feasible frontier. The hour of work that he does by giving up an hour of free time gets Omar $50, and each dollar gets him 4km, so the MRT is 200: giving up an hour of free time gets him 200km of feasible air travel.

Omar's preferences for free time and air travel are given by the indifference curves shown. The slope of the indifference curve indicates how much he values free time relative to air travel, that is, his MRS of free time for air travel.

We can see that the highest indifference curve that Omar can reach (at point *A*) results from his choosing to work 200 extra hours so as to have 6,760 hours of free time and 40,000km of air travel.

To Omar, the private cost of a mile of air travel is $0.25. But we know that the social costs—the private costs plus the costs of the emissions due to burning aviation fuel and other external effects—are not included in his private cost calculation. Now imagine that a policy is adopted with the objective of inducing Omar to internalise the full social cost of his vacation choices, by raising the price of air travel so that the private cost to Omar is equal to the social cost. A tax is levied on aviation fuel, so a dollar spent on a ticket now purchases only 2km. The new feasible frontier and feasible set is shown in the figure as the dashed line. The new marginal rate of transformation is 100km of travel per hour of free time given up.

How will the tax affect Omar's decision? As before, Omar chooses the point on the feasible frontier that is on the highest indifference curve, which is now point B. He flies less. There are two reasons for the change:

- *The income effect:* Omar is less well-off than before because the price of something that he consumes has gone up. His real income has fallen.
- *The substitution effect:* The tax has increased the relative price of air travel, leading Omar to substitute other ways of having a good life, by consuming other goods, possibly by working less, or both.

18.9 WHY IS ADDRESSING CLIMATE CHANGE SO DIFFICULT?

While scientists agree that climate change is occurring and that our economic activity is contributing to it, there are large gaps in scientific understanding of the processes involved and the costs of containing them.

Moreover, as we have seen in sections 18.4 and 18.5, conflicts of interest over the extent and methods of abatement make it difficult for national governments to adopt broadly supported strategies for mitigating environmental degradation. These conflicts often take the form of disagreements about what climate science has shown. In the United States in 2015, 64% of Democratic Party supporters were of the opinion that global warming both is occurring and is a result of human activity. The similar fraction among Republicans was 22%.

Owners and employees of companies producing or using fossil fuels anticipate income losses as the result of policies to reduce emissions, and spend heavily to influence public opinion on environmental questions. You can read about the impact of this here and here.

Partly as a result, few citizens around the world place a higher value on environmental problems than on the economy, as shown in Figure 18.23.

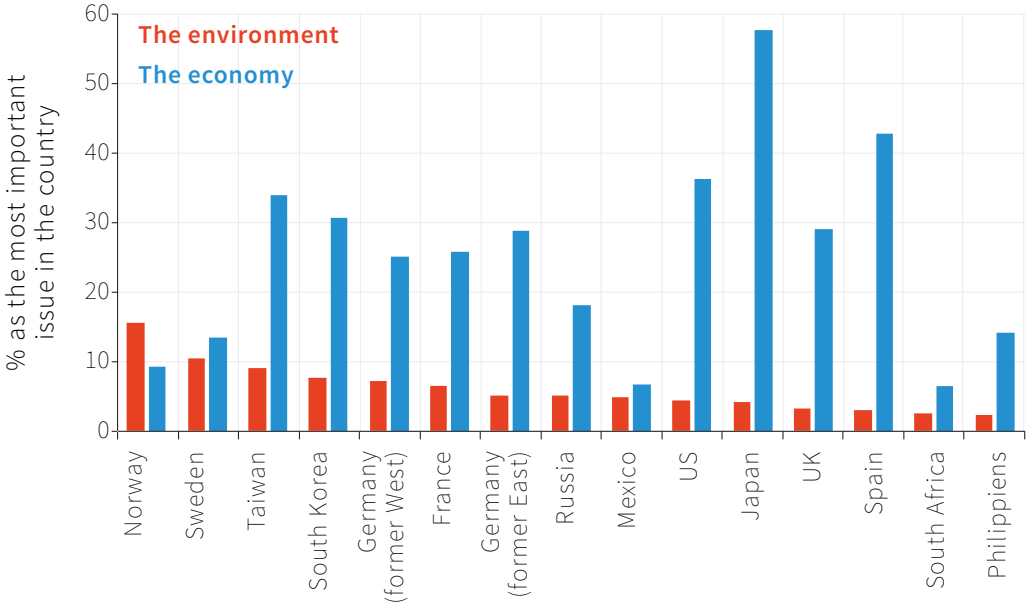


Figure 18.23 *Survey views on the importance of the environment and the economy as an issue.*

Source: *ISSP Research Group. 2012. 'International Social Survey Programme: Environment III - ISSP 2010', August. GESIS Data Archive, Cologne. Note: The question asked was "Which of these issues is the most important for [COUNTRY] today?"*

Lack of adequate information and conflicts of interest are impediments to good public policy in many areas, not just climate change. But addressing climate change faces two unusual challenges: the problem cannot be solved by national governments acting alone, and those affected by our choices today include generations in the distant future.

International cooperation

Using the tools of game theory in Unit 4, we saw that avoiding the *tragedy of the commons* that afflicts the supply of public goods depends on the rules of the game (the institutions). Where there are repeated interactions of the players and there are opportunities to punish those who do not contribute to the public good, the socially optimal outcome can be sustained. The presence in several continents of sustainable water-use systems or fish stocks shows that the tragedy of the commons is not inevitable.

In the case of climate change, game theory helps us understand the obstacles to its solution. Recall the way we modelled the climate change game as a prisoners' dilemma in which two countries (the US and China) can either restrict carbon emissions or continue with business as usual (see Figure 4.17). Self-interest makes the business as usual scenario the dominant strategy equilibrium.

To understand how an international agreement might be negotiated to avoid the business as usual outcome, we introduced inequality aversion and reciprocity. If citizens of the US and China give some weight to the wellbeing of citizens in the other

country or experience less wellbeing when inequality rises, and if they are willing to implement costly measures as long as this is done in the other country, then an outcome where both countries restrict emissions is possible.

Our hypothetical model of climate change negotiations between China and the US gave rise to two Nash equilibria if citizens had both inequality aversion and reciprocity. It is not completely unrealistic: after intense negotiations following failed talks and a non-binding agreement in Copenhagen in 2009, *all* countries committed to eventual emission cuts at the United Nations Conference on Climate Change in Paris in December 2015 with the goal of stabilising global temperatures at 2C above pre-industrial levels. Virtually all countries also submitted their individual plans for cutting emissions, but these plans are not yet consistent with this temperature stabilisation goal.

Unrepresented generations

Our economic activity today will be felt in climate changes in the distant future. So we are essentially creating consequences that others will bear. This is just an extreme form of external effects that we have studied throughout the course. It is extreme not only in its potential consequences, but also in that those who will suffer the consequences are future generations.

In many countries public policies have been adopted to address other kinds of environmental external effects—such as local pollution—under pressure from voters bearing the costs of these effects. If you look ahead at Figure 18.25b, you will notice that many of the stars (well above the line) on the Environmental Performance Index are, and have long been, electoral democracies. This is not the case for most of the low performers.

The future generations that will bear the consequences of our decisions are unrepresented in the policymaking process today. The only way the wellbeing of these unrepresented generations will be taken into account at the environmental bargaining tables around the world is the fact that—as we have seen in Unit 4—people (at least most of us, some of the time) care about, and would like to behave ethically toward, others.

This is what lies behind the debates among economists about how much we should value the future benefits and costs of the decisions about climate that we make today.

DISCOUNTING FUTURE GENERATIONS' COSTS AND BENEFITS

A measure of how we value today the benefits of our actions to other people who will live in the future.

- Note this is *not* a measure of individual *impatience* about one's own future benefits and costs

In considering alternative environmental policies, how much we value the wellbeing of future generations is commonly measured by an interest rate: it is literally the rate at which we discount (literally, count less) future people's costs or benefits. There are, however, debates about how this discounting process should be done.

WHEN ECONOMISTS DISAGREE

THE DISCOUNTING DILEMMA: HOW SHOULD WE ACCOUNT FOR FUTURE COSTS AND BENEFITS?

When considering policies, economists seek to compare the benefits and costs of alternative approaches. Doing this presents especially great challenges when the policy problem is climate change. The reason is that the costs will be borne by the present generation but the benefits of a successful abatement policy will be enjoyed by people in the future, many of them not yet alive.

Put yourself in the shoes of the impartial policymaker we studied earlier and ask yourself: are there any reasons why, in summing up the benefits and costs of an abatement policy, I should value the benefits expected to be received by future generations any less than the benefits and costs that will be borne by people today? Two reasons come to mind:

- *Technological progress:* The people in the future may have either greater or lesser needs than we do today. For example, as a result of continuing technical improvements, they may be richer (either in goods or free time) than we are today, so it might seem fair that we should not value the benefits they will receive from our policies as highly as we value the costs that we will bear as a result.
- *Extinction of the human species:* There is a small possibility that the future generations will not exist because humanity becomes extinct.

These are good reasons why we might discount the benefits received by future generations. Notice that neither of these reasons for discounting is related to impatience.

This was the approach adopted in the 2006 *Stern Review on the Economics of Climate Change* (read the executive summary here). Nicholas Stern, an economist, selected a *discount rate* to take account of the likelihood that people in the future would be richer: based on an estimate of future productivity increases, Stern discounted the benefits to future generations by 1.3% per annum. To this he added a 0.1% per annum discount rate to account for the risk that in any future year there might no longer be surviving generations.

Based on this assessment, Stern advocated policies that would have implemented substantial abatement investments today to protect the environment of the future.

Several economists, including William Nordhaus, criticised the Stern Review for its low discount rate. Nordhaus wrote that Stern's choice of discount rate "magnifies impacts in the distant future". He concluded that, with a higher discount rate, "the Review's dramatic results disappear".

Nordhaus advocated the use of a discount rate of 4.3%. (The next box illustrates what a big difference from Stern's number this really is.) Discounting at this rate means that a $100 benefit occurring 100 years from now is worth $1.48 today. At Stern's 1.4% rate it would be worth $24.90. This means a policymaker using Nordhaus' discount rate would approve of a project that would save future generations $100 in environmental damages if it cost less than $1.48 today. A policymaker using Stern's 1.4% would approve the project if it cost less than $24.90.

Not surprisingly, then, Nordhaus' recommendations for climate change abatement were far less extensive and less costly than those proposed by Stern. To deter the use of fossil fuels, for example, Nordhaus advocated a carbon price of $35 per tonne in 2015. Stern recommended a price of $360.

Why did the two economists differ by so much? They agreed on the need to discount for the likelihood that future generations would be better off. But Nordhaus had an additional reason to discount future benefits: impatience.

Reasoning as we did in Unit 11 for Julia's and Marco's consumption now or later, Nordhaus used estimates based on market interest rates as measures of how people today value future versus present consumption. Using this method he came up with a discount rate of 3% to measure the way people discount future benefits and costs that they themselves may experience. Nordhaus included this in his discount rate, which is why Nordhaus' discount rate (4.3%) is so much higher than Stern's (1.4%).

Critics of Nordhaus pointed out that in evaluating the claims that future generations should have on our concern, a psychological fact like our own impatience is not a reason to discount the needs and aspirations of other people in future generations.

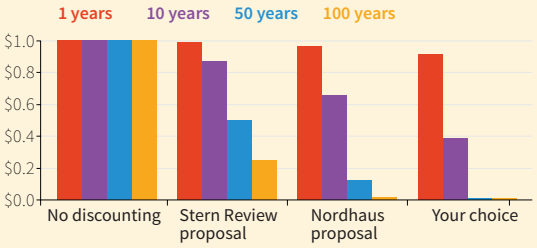
Stern's approach counts all generations as equally worthy of our concern for their wellbeing. Nordhaus, in contrast, takes the current generation's point of view and counts future generations as less worthy of our concern than the current generation, much in the way that, for reasons of impatience, we typically value current consumption more highly than our own future consumption.

Is the debate resolved? The discounting question ultimately requires adjudicating between the competing claims of different individuals at different points of time. This involves questions of ethics on which economists will continue to disagree.

DISCUSS 18.8: SIMULATING DIFFERENT DISCOUNT RATES

Download the simple discount rate simulation spreadsheet (right) from our website. The simulator allows you to calculate the present value of receiving $1 in one, 10, 50 and 100 years from now for four discount rates.

Discount rate (%)	Source	0	1	10	50	100
		Years in the future				
0.0	No discounting	$ 1.00	$ 1.00	$ 1.00	$ 1.00	$ 1.00
1.4	Stern Review proposal	$1.00	$0.99	$0.87	$0.50	$0.25
4.3	Nordhaus proposal	$1.00	$0.96	$0.66	$0.12	$0.01
3.0	Your choice	$1.00	$0.97	$0.74	$0.23	$0.05



In the spreadsheet, the first three discount rates are fixed: zero, Stern's suggestion, and the alternative suggested by Nordhaus.

1. Explain the effect of different discount rates on the present value of receiving $1 in the future.

The fourth rate is your choice: use the slider in the table to choose a discount rate you think is appropriate for the evaluation of climate change policy.

2. Justify your choice. Is it closer to the Nordhaus or Stern proposal?
3. Try to find out what discount rate your government (or another government of your choice) uses to evaluate public investment projects. Do you think it is appropriate?

18.10 POLICY DEBATES

We have introduced price-based and quantity-based policies. They may affect the environment both in a static way (moving to or along a given feasible frontier with given indifference curves) or a dynamic way (changing technologies and, in the long run, values).

We summarise these distinctions and give examples in Figure 18.24:

	PRICE	QUANTITY
STATIC	A carbon tax increases the incentive for households and firms to choose an alternative energy source.	Ban on lead in petrol (US 1996; China 2000) facilitated the use of more environment-friendly engines, and eliminated a health hazard.
DYNAMIC	A carbon tax would increase the profits of innovators in nuclear and wind, solar and other renewable energy sources.	Ban on ozone-depleting substances (for example CFCs in Montreal Protocol 1987) stimulated development of alternative technologies.

Figure 18.24 *Addressing environmental external effects.*

Differences between countries

Environmental policies make a difference. We can see countries vary greatly in the global environmental damage they inflict and in their success at managing environmental quality in their country. Figure 18.25a shows CO_2 emissions per capita for each country in 2010 alongside the income per capita. Richer countries produce more CO_2 per capita than poorer ones. This is to be expected because greater income per capita is the result of a higher level of production of goods and services per capita, with associated impacts on the biosphere. This is what the upward-sloped line indicating the relationship between the two variables shows.

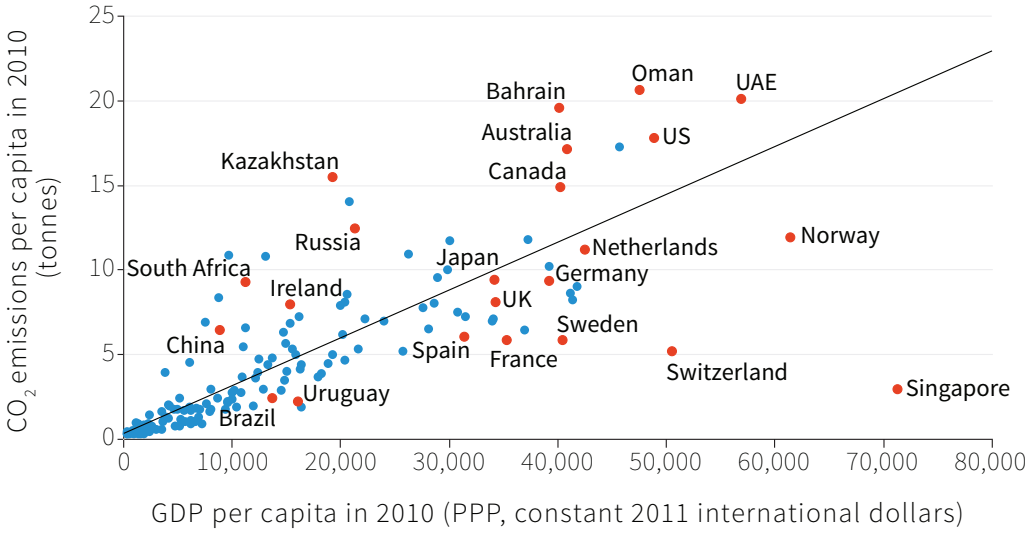


Figure 18.25a *Carbon dioxide emissions are higher in richer countries…*

Source: *The World Bank. 2015. 'World Development Indicators.' Three small very high-income countries—Kuwait, Luxembourg and Qatar—are not shown.*

But notice, too, that among countries at approximately the same level of per capita income, some emit much more than others. Compare the high emissions levels in the US, Canada, and Australia with the lower emissions levels of France, Sweden and Germany, countries at approximately the same level of per-capita income. Another

way to read the graph is horizontally: Norway has the same emissions level that would be predicted (by the line) for a country $20,000 poorer in per capita income. Russia pollutes as much as would be expected from a country $20,000 richer.

Singapore is an high-performing outlier. It is a high-income city-state with an effective public transport network and a commercial rather than industrial economic base, resulting in limited levels of pollution. In addition to public transportation, the government has adopted other effective environmental policies. For example, if you want to use a car in Singapore, you are first required to purchase a permit for a car at an auction, and then pay the congestion charge (a tax) every time you drive into the city.

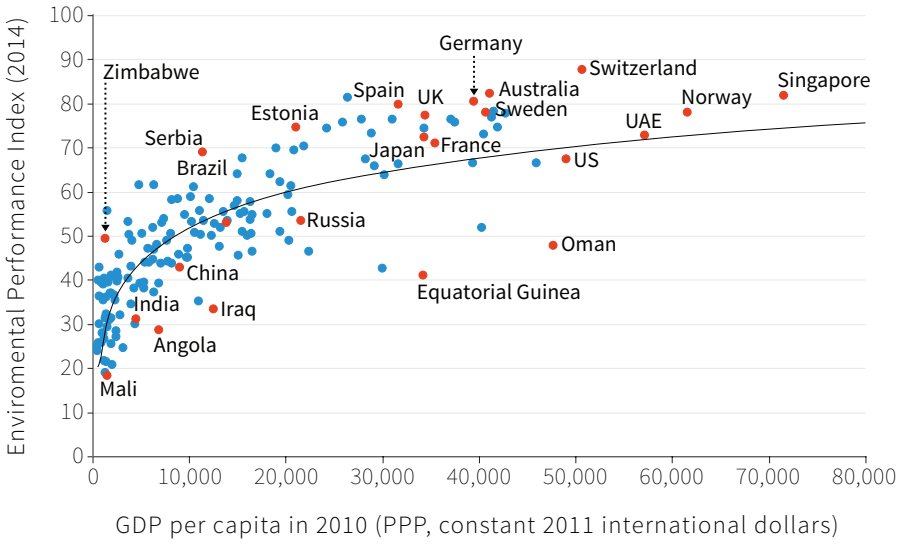


Figure 18.25b *...but so is the quality of the local environment.*

Source: *Development Indicators; EPI. 2014. 'Environmental Protection Index 2014.' Yale Center for Environmental Law & Policy (YCELP) and the Center for International Earth Science Information Network.*

Though richer countries emit more CO_2 per capita, they have also adopted more effective policies to manage their own environmental resources, such as forests, soil, biodiversity and water. Figure 18.25b plots the Environmental Performance Index (EPI) against GDP per capita. The EPI is a broad index of country-level environmental health and ecosystem vitality, including the state of wastewater treatment, fisheries and forests. It brings together 20 different country-level indicators including trends in carbon emissions, fish stocks, changes in forest cover, quality of wastewater treatment, access to sanitation, air pollution and child mortality. In this case a curved rather than straight line fits the data better, indicating that differences in per capita income are associated with major differences in the EPI for very poor countries, but not as major for the richer countries, on average.

As in the previous figure, Russia underperforms, with the environmental performance index expected of a country half as rich. Germany, Sweden and Switzerland are high performers. Notice that Australia, which is an unusually big

emitter of CO_2 (Figure 18.25a), is a top performer on the national environmental amenities measured by the EPI. A good part of the environmental damage done by economic activity in Australia is thus imposed as a cost on those outside the country.

The message of this figure is similar to the previous one: countries—even at similar levels of income per capita—differ greatly in their environmental performance. Compare Switzerland with the US or Spain with Russia, for example. Both India and China are substantially below the line. These country differences suggest the importance of the kinds of policies that are adopted and enforced.

DISCUSS 18.9: HIGH AND LOW PERFORMERS

Consider the labelled countries above the best-fit line in Figure 18.25b and those below the line.

1. What facts about the countries do you think might explain their status as high and low performers respectively?
2. Find out about environmental policies and political systems of these countries using The World Bank Development Indicators and Freedom in The World 2016. What information from these sources helps you to explain the differences between high and low performers, and how does it help?

Evaluating market-based policies

Market-based policies make use of the information often not available to governments but which is contained in prices that (when adjusted by environmental taxes and subsidies) ideally reflect the marginal costs and benefits that should be taken into account when a firm or individual is considering an action with external environmental effects.

But as in the case of housing and financial assets, the environmentally relevant prices often diverge considerably from this ideal, as we have seen in the collapse in the price of carbon emissions permits after the financial crisis.

Among the market-based policies, taxes and the sale of permits can raise significant amounts of government revenue that can then either fund socially valuable projects or allow the elimination of sources of revenue—taxes that discourage employers to hire, or invest—that impose deadweight losses on the economy.

The case for market-based policies is typically made by reference to an equilibrium in which the relevant actors have exploited all gains; but as we have seen in Unit 9 the state of the economy is often far from an equilibrium of this sort. In Figure 18.26 we look again at the estimates of the marginal abatement costs that we previously saw

in Figure 18.7 (note that we have rotated the axes 90 degrees clockwise so that we can fit new information on it). In Figure 18.7 we included only costly policies that could be promoted as an objective of government policy. Figure 18.26 additionally includes actions that would accomplish significant abatement, *and would also have monetary benefits greater than the costs*. In the figure, when the monetary benefit is greater than the cost, the bar extends to the left. When cost is greater, it extends to the right.

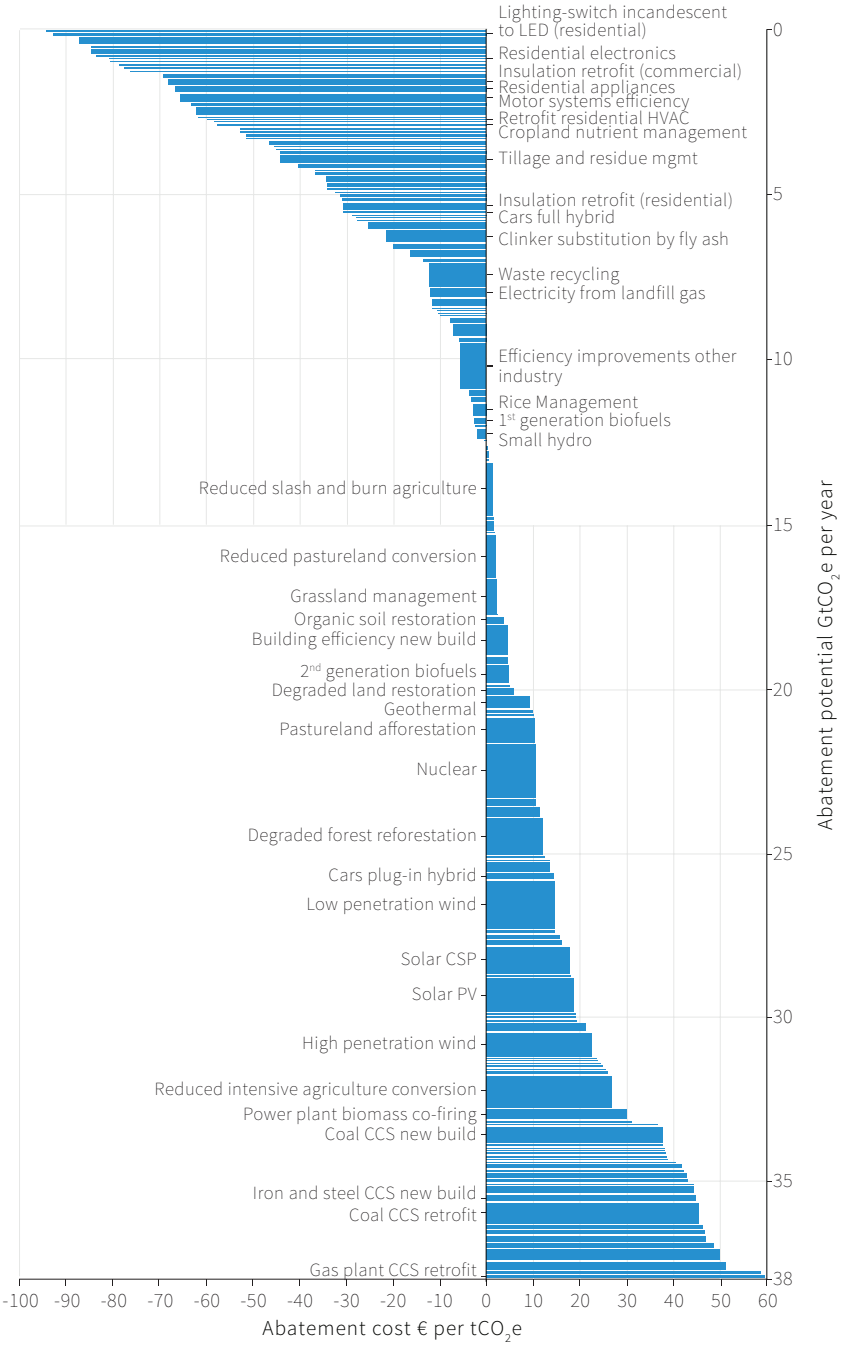
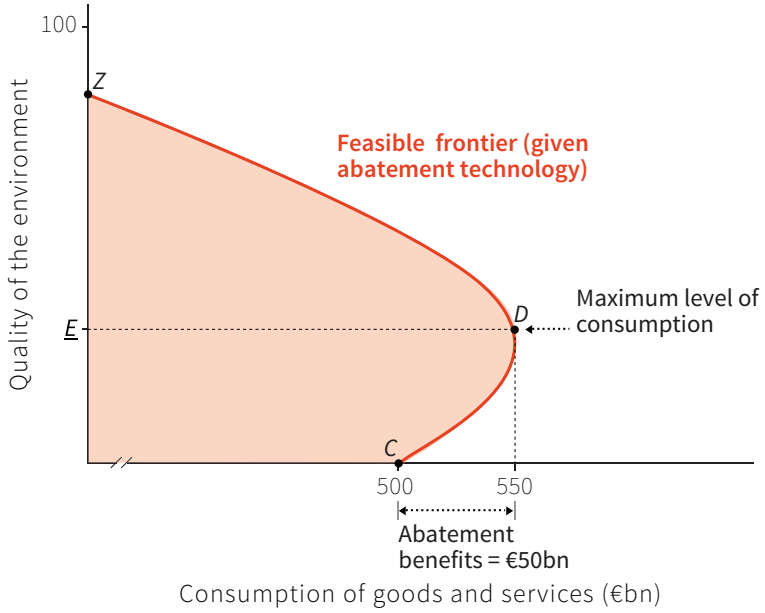


Figure 18.26 *Global greenhouse gas abatement curve: Abatement in 2030 as compared with business as usual.*

Source: *McKinsey & Company. 2013. Pathways to a Low-Carbon Economy: Version 2 of the Global Greenhouse Gas Abatement Cost Curve. McKinsey & Company.*

Replacing incandescent light bulbs by LED bulbs in our houses would be the most cost-saving policy of all but it is a narrow bar, meaning it does not have a big abatement potential. Fuel-efficient vehicles, insulation in houses and offices, and other technologies with bars to the left are also cost-saving. Note that if we were to adopt only cost-saving policies between now and 2030, we would still achieve more than a quarter of the total potential abatement.

We can represent the unrealised abatement potential of these changes in the feasible set figure. Start at point C on the horizontal axis in Figure 18.27. The evidence from Figure 18.26 is that implementing the measures (starting on the left in Figure 18.26, with replacement of incandescent bulbs by LEDs) will generate abatement benefits *and* at the same time allow for higher consumption of other goods and services. This produces the positively sloped part of the feasible frontier, with both environmental quality and consumption rising from *C* to *D*. Once all the measures have been introduced that reduce costs, at *D*, it begins to be costly to achieve further abatement and the feasible frontier is negatively-sloped, as we saw when we analysed the implications of Figure 18.7. Point *D* corresponds to the point of maximum consumption and zero abatement (at €500bn) that we saw in Figure 18.12.



Implementing abatement along the feasible frontier

Moving from *D* to *Z* takes the quality of the environment above *E* but at the cost of lower consumption.

Figure 18.27 *Is there always a trade-off between consumption and environmental quality?*

The unrealised abatement potential of these changes, despite the fact that they would save money for the individuals or firms implementing them, suggests that implementation by market incentives may be slow and incomplete. There are two responses to this:

- We could try to understand why people do not take environment-friendly actions even when they are cost-reducing.
- Complement market-based policies with quantity-based policies.

Evaluating quantity-based policies

A primary advantage of quantity-based policies, when the government has the necessary information and enforcement capacities, is that implementation can be rapid and complete. An example is the dramatic reduction in the use of lead in petrol in many countries around the world following a ban.

The information necessary for a government to enforce a ban is typically far less than that required to implement a tax and subsidy policy.

Quantity measures, in isolation, do not make use of the valuable (if not ideal) information that private economic actors reveal through the prices at which they are willing to transact.

Fairness

It is widely accepted that fairness is an important standard to judge outcomes, though some economists consider these judgements to lie outside of economics. The controversy surrounds value judgements of fairness like those often associated with the polluter pays principle.

This principle can be interpreted as an application of the basic economics of environmental policies. Environmental external effects often impose costs on others, and making the polluter pay for these external effects is a way to internalise (and therefore eliminate) them.

This could be accomplished by taxing the polluting activity so as to raise the private marginal cost to correspond to the marginal social cost, as was shown in Figures 18.15a and 18.15b. This may be an efficient way to abate the pollution. But notice from those figures that the same abatement could be accomplished by providing the firm with a subsidy for the use of an alternative technology that resulted in a lower level of emissions.

The firm's-eye view of these two policies may be that the tax is the stick and the subsidy the carrot. The tax, which reflects the polluter pays principle, lowers the profits of the firm. A subsidy raises the firm's profits. Whether the carrot or the stick is the right policy depends on such things as:

- The feasibility and cost of the implementing the subsidy compared to the tax.

- Reasons (not necessarily stemming from environmental concerns) that a policymaker would want to raise or lower the firm's profits in this way.

Examples of reasons for changing the firm's profits include a desire to provide incentives for the firm to invest or a concern for fairness, motivating policies to redistribute income from those who receive profits to those are less well off.

The polluter pays principle is not always a good guide to the best policy. Think of a large city in a low-income country in which much of the cooking is still done over wood fires, generating high levels of airborne particulate matter and causing asthma and other respiratory illnesses.

- *Fairness:* It is mostly poor families who lack the income or access to electricity that would allow them to cook and heat their homes with fewer external environmental effects. Many would object in this case on fairness grounds to making the polluters pay, and instead favour subsidising kerosene or providing a better electricity supply.
- *Effectiveness:* Subsidising kerosene is likely to be cost-effective in reducing smog compared to tracking down and extracting payments from hundreds of thousands of people who are polluting the city's air with wood fires.

This example is helpful because it shows not only the value of considering fairness as well as efficiency, but also the importance of being clear about which objective we are pursuing when we design policies.

18.11 CONCLUSION

For 100,000 years or more, humans—like other animals—lived in ways that modified the biosphere but did not substantially and irreversibly degrade its capacity to support life on the planet. Starting 200 years ago, humans learned how to use the energy available from nature to transform how we produced goods and services, radically increasing the productivity of our labour. The capitalist economy provided both the carrots and the sticks that made the technological revolution profitable to private firms and hence a permanent feature of our lives. The result was a sustained increase in the output of goods and services per person.

In many countries the extension of the vote to people who worked as employees and their organisation into trade unions and political parties enhanced the bargaining power and the wages of workers. The increasing cost of hiring labour provided

particular incentives for owners of firms to seek innovations that would use less labour, substituting machinery and the non-human energy of coal and other fuels that powered them for labour.

The result of this process—increased productivity and bargaining power of labour—was in many countries the growing affluence of workers. But the substitution of non-human energy to power the machines for human labour also led to the impoverishment of nature.

The impoverishment of nature cannot be reversed, however, by the same mechanism that created this affluence. Workers were their own advocates, and their success in pursuing their private interests in seeking a higher living standard led to the wage increases, resulting in a pattern of technological change in which less labour was used in production. Future generations and non-human elements of the contemporary biosphere are not capable of advocating for saving nature the way workers indirectly advocated for saving labour.

The imposition of prices on the use of nature sufficient to deter the degrading external effects of the production of goods and services today will require public policies as well as private bargaining. Should this occur, it will be propelled not by the silent voices of the biosphere and generations unborn, but by people today, concerned not primarily about their private interests, but about the preservation of a flourishing biosphere in the future.

CONCEPTS INTRODUCED IN UNIT 18

Before you move on, review these definitions:

- *Abatement*
- *Abatement policies*
- Natural *resources* and *reserves*
- *Global greenhouse gas abatement cost curve*
- *Environment-consumption indifference curve*
- *Marginal productivity* and *opportunity cost* of abatement expenditures
- *Price-* and *quantity-based* environmental policies
- *Cap and trade*
- *Contingent valuation*
- *Hedonic pricing*
- *Discounting future generations' costs and benefits*
- The *polluter pays principle*
- *Tipping point*
- *Austerity policy*

Key points in Unit 18

The biosphere

The economy is part of the Earth's biosphere, which has limited capacity to sustain a growing economy that relies on fossil fuels.

How much abatement?

The extent to which environmental damages should be abated depends on both the costs of abatement and the benefits of a sustainable environment relative to other valued objectives.

Costs and benefits of abatement

These costs and benefits are summarised in the marginal rate of transformation of foregone consumption into environmental quality (based on the marginal abatement cost curve) and the marginal rate of substitution between consumption and environmental quality.

Conflicts of interest

Conflicts over the extent and methods of abatement arise because different people do not share equally the costs or the benefits of a less degraded environment.

The polluter pays

A major objective of environmental policy is that consumption or production activities that degrade the environment should bear the environmental costs, so that the prices that affect our decisions more closely approximate the marginal social costs (including environmental external effects).

Abatement policies

Policies that accomplish this objective when addressing climate change include carbon taxes and tradable carbon emissions permits.

Future lifestyles

These policies also will promote low-carbon technologies and lifestyles in the future.

Putting a price on the environment

Economists measure the costs of a degraded environment using contingent valuation and hedonic pricing. A shortcoming of both is that the preferences of those with less wealth are counted less than those of the better off. Others consider a healthy environment to be a merit good.

Future generations

Economists do not agree on how best to value the environmental benefits of future generations.

18.12 EINSTEIN

Marginal abatement costs and the total productivity of abatement expenditures

How do we construct the line segments that define the boundary of the feasible set in Figure 18.8 from the data in Figure 18.7?

Let the height of the first bar (the most cost-effective abatement expenditure) in Figure 18.7 be y and the width of that bar be x. Then, in Figure 18.8:

- The initial slope of the curve is $1/y$
- The horizontal axis value of the first point is xy
- This point's vertical axis value is x

The other line segments making up the curve in Figure 18.8 are constructed in the same way.

Environment-consumption indifference curves and the marginal rate of substitution

In Figure 18.12, suppose that each citizen places a value of 1 on each unit of consumption that he or she can enjoy, and a value of μ on the quality of the environment. The quality of the environment is E and the amount each citizen can consume (C) is total income (Y) minus total abatement costs (A) divided by the total population (n). So the citizens' utility (u) is:

$$\begin{aligned} u &= \mu E + C \\ &= \mu E + \frac{(Y - A)}{n} \end{aligned}$$

In other words, the utility of an individual is the value placed on environmental quality multiplied by the quality of the environment (in units of abatement achieved), plus the consumption of goods and services per person.

This equation makes clear the difference between the public good (E) and the private good (C): the former is something that is consumed by everyone, the latter is divided up among the members of the population.

Indifference curves based on this utility function have slopes equal to the ratio of the marginal utility of a better environment μ, to the marginal disutility of a greater total societal expenditure on the environment which is the fraction of the expenditure the citizen will pay ($1/n$), multiplied by the value of the consumption that will be forgone if more is spent on abatement, which is 1.

Thus the marginal rate of substitution (the slope of the indifference curve) is:

$$MRS = \frac{\textit{marginal disutility of abatement spending}}{\textit{marginal utility of improved environment}}$$
$$= \frac{1/n}{\mu}$$

Indifference curves when the costs of abatement are not equally shared

In Figure 18.13, assume businesses pay a fraction β of the costs of abatement and citizens pay 1- β. For simplicity, think about a population with just two individuals: a citizen and a business. The polluter pays principle means that $\beta > 0.5$. The citizen gets a share w of the total income (the business gets the remainder), so her utility function is:

$$u^c = \mu E + wY - (1 - \beta)A$$

In other words, the citizen's utility is the utility from the quality of the environment, plus wage income, minus her contribution to cost of abatement.

The business has utility:

$$u^b = \mu E + (1 - w)Y - \beta A$$

Which means that the business has utility equal to the utility from the quality of the environment (the same as citizen's), plus profit income, minus the contribution to cost of abatement (which is greater than the citizen's).

To allow us to concentrate on the implications of conflicts of interest over who pays for abatement, we assume that both the citizen and the business care equally about the environment. This is represented in the model by μE.

Then the marginal utility of environmental quality is μ for both of them. They differ only in who bears the cost of abatement. The marginal disutility of abatement expenditures for the citizen is now (1 - β), which is less than that for the business (β). This shows up when we compare the slopes of the indifference curves:

$$MRS = \frac{\textit{marginal disutility of abatement spending}}{\textit{marginal utility of environmental improvement}}$$

For the business:

$$MRS^b = \frac{\beta}{\mu}$$

For the citizen:

$$MRS^c = \frac{(1 - \beta)}{\mu}$$

Because $\beta > 1/2$ (the business pays more of the cost of abatement), we know that :

$$\frac{\beta}{\mu} > \frac{1/2}{\mu} > \frac{(1-\beta)}{\mu}$$

This means that the indifference curve for the business is steeper (less "green") than for the citizen.

READ MORE

Bibliography

1. Allcott, Hunt, and Sendhil Mullainathan. 2010. 'Behavior and Energy Policy.' *Science* 327 (5970): 1204–5.
2. Aurora Energy Research. 2014. 'Carbon Content of Global Reserves and Resources.'
3. BP Global. 2015. 'Statistical Review: Energy Economics.'
4. Boyce, James K. 2012. 'Is Inequality Bad for the Environment?' In *Economics, the Environment and Our Common Wealth*, by James K Boyce. Cheltenham: Edward Elgar Publishing.
5. Bundesanstalt für Geowissenschaften und Rohstoffe (The Federal Institute for Geosciences and Natural Resources). 2012. *Energy Study 2012*.
6. Burtraw, Dallas. 2000. 'Innovation Under the Tradable Sulfur Dioxide Emission Permits Program in the U.S. Electricity Sector.' *Resources for the Future* Discussion Paper 00–38.
7. Chen, Yuyu, Avraham Ebenstein, Michael Greenstone, and Li Hongbin. 2013. 'Evidence on the Impact of Sustained Exposure to Air Pollution on Life Expectancy from China's Huai River Policy.' *Proceedings of the National Academy of Sciences* 110 (32): 12936–41.
8. EPI. 2014. 'Environmental Protection Index 2014.' Yale Center for Environmental Law & Policy (YCELP) and the Center for International Earth Science Information Network.
9. Etheridge, D.M., L.P. Steele, R.L. Langenfelds, R.J. Francey, J-M. Barnola, and V.I. Morgan. 1996. 'Natural and Anthropogenic Changes in Atmospheric CO2 over the Last 1000 Years from Air in Antarctic Ice and Firn.' *Journal of Geophysical Research* 101: 4115–28.
10. Freedom House. 2016. *Freedom in the World 2016*.
11. Gillingham, Kenneth. 2014. 'Identifying the Elasticity of Driving: Evidence from a Gasoline Price Shock in California.' *Regional Science and Urban Economics* 47 (July): 13–24.
12. Goulder, Lawrence H, and Roberton C Williams. 2012. 'The Choice of Discount Rate for Climate Change Simulation.' *Climate Change Economics* 03 (04): 1250024.

13. Hepburn, Cameron, Eric Beinhocker, J. Doyne Farmer, and Alexander Teytelboym. 2014. 'Resilient and Inclusive Prosperity within Planetary Boundaries.' *China & World Economy* 22 (5): 76–92.
14. IPCC. 2013. *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge: Cambridge University Press.
15. ISSP Research Group. 2012. 'International Social Survey Programme: Environment III - ISSP 2010', August. GESIS Data Archive, Cologne.
16. Kristof, Nicholas D. 2013. 'This Is Your Brain on Toxins.' *The New York Times*, October 16.
17. Lazard. 2015. 'Levelized Cost of Energy Analysis 9.0.' Lazard.com. November 17.
18. McKinsey & Company. 2013. *Pathways to a Low-Carbon Economy: Version 2 of the Global Greenhouse Gas Abatement Cost Curve*. McKinsey & Company.
19. Millennium Ecosystem Assessment. 2005. *Ecosystems and Human Well-Being: Synthesis*. Washington, DC: Island Press.
20. Nagy, Béla, J. Doyne Farmer, Quan M Bui, and Jessika E Trancik. 2013. 'Statistical Basis for Predicting Technological Progress.' *PLoS ONE* 8 (2). Public Library of Science.
21. Nemet, Gregory F. 2006. 'Beyond the Learning Curve: Factors Influencing Cost Reductions in Photovoltaics.' *Energy Policy* 34 (17): 3218–32.
22. Nordhaus, William D. 2007. 'A Review of the Stern Review on the Economics of Climate Change.' *Journal of Economic Literature* 45 (3): 686–702.
23. OpenSecrets.org. 2015. 'Lobbying Spending Database Chemical & Related Manufacturing.'
24. Porter, Michael E, and Claas van der Linde. 1995. 'Toward a New Conception of the Environment-Competitiveness Relationship.' *Journal of Economic Perspectives* 9 (4): 97–118.
25. Schmalensee, Richard, and Robert N Stavins. 2013. 'The SO2 Allowance Trading System: The Ironic History of a Grand Policy Experiment.' *Journal of Economic Perspectives* 27 (1): 103–22.
26. Skånberg, Kristian. 2001. *Constructing a Partially Environmentally Adjusted Net Domestic Product for Sweden 1993 and 1997*. UN Committee of Experts on Environmental-Economic Accounting (UNCEEA).
27. Smith, Stephen. 2011. *Environmental Economics: A Very Short Introduction*. Oxford: Oxford University Press.
28. Stavins, Robert N, Gabriel Chan, Robert Stowe, and Richard Sweeney. 2012. 'The US Sulphur Dioxide Cap and Trade Programme and Lessons for Climate Policy.' *VoxEU.org*. August 12.
29. Stern, Nicholas. 2007. *The Economics of Climate Change: Stern Review on the Economics of Climate Change*. Cambridge: Cambridge University Press.
30. Stokes, Bruce, Richard Wike, and Jill Carle. 2015. *Global Concern about Climate Change, Broad Support for Limiting Emissions*. Pew Research Center's Global Attitudes Project.

31. The World Bank. 2011. 'The Changing Wealth of Nations Database.'
32. The World Bank. 2015. 'Commodity Price Data.'
33. The World Bank. 2015. 'World Development Indicators.'
34. Wagner, Gernot, and Martin L. Weitzman. 2015. *Climate Shock: The Economic Consequences of a Hotter Planet*. Princeton, NJ: Princeton University Press.

19 ECONOMIC INEQUALITY

(c) Tuca Vieira

ECONOMIC DISPARITIES ARE MOSTLY A MATTER OF WHO YOUR PARENTS ARE; WELL-DESIGNED POLICIES AND INSTITUTIONS CAN REDUCE INEQUALITIES WITHOUT LOWERING AVERAGE LIVING STANDARDS

- Your income depends on who your parents are because this will determine the country in which you are born, your ethnic group and your first language. It will also influence the amount and quality of your education, inherited wealth and social networks
- Our model of the labour market helps explain the way that policy choices and new technologies alter inequality as measured by the Gini coefficient
- Income inequality declined within most countries during most of the 20th century, a trend that has reversed in many of them since the 1980s
- Income inequality among the citizens of the world increased between the early 19th century and the end of the 20th century, and has declined since because of the rapid economic growth of China and India
- Economic inequalities arising from discrimination based on race, gender, or religion are considered by many people to be unfair, as are inequalities arising from other forms of unequal opportunity
- While inequalities may provide incentives for hard work and risk-taking, they may also bring with them costs that impair economic performance
- The trade-offs that a society faces in addressing inequality can be modelled using the standard framework of feasible sets and indifference curves
- Well-designed and implemented government policies can limit economic inequality without reducing average living standards

Institute for New Economic Thinking
UCL SciencesPo Azim Premji University
FRIENDS PROVIDENT Foundation

It is 1975. On a collective farm in China, Yichen Li and Renfu Bo, both 17, live under Communist Party rule. Renfu is the child of a local Communist Party leader who is close to Deng Xiaoping, the soon-to-be leader of China. In 10 years he will attend Tsinghua University, an elite engineering university in Beijing, and will join the Communist party himself. In 20 years he will run a state-owned enterprise. In 30 years he will be CEO of the company after it is privatised, and be highly ranked within the Party.

Yichen, however, will not go to university, but instead will work the land alongside her parents—who have no party connections—until she is 16, then will try working at a state-owned enterprise making car parts for export to the US and Europe. When she is 30 years old she will hear about an opportunity to work in the new Motorola factory opening in nearby Guangdong, paying three times her current wage. She will decide to go there, but will not be able to obtain a *hukou* (internal passport) to migrate legally to Guangdong. She will travel anyway, giving up access to health care and to public education for her daughter. She will live in barracks with the other workers, missing her family.

Yichen and Renfu are hypothetical. We could have inserted a disclaimer: "All characters appearing in this work are fictitious...". But that would not be entirely true—they illustrate the divergent real life histories of real people alive today.

Let's also consider two other hypothetical people, living in the US, also in 1975. Mark and Stephanie, also 17, live in Gary, Indiana. Mark is about to finish high school and start working in the local unionised steel mill with his father, where the pay is good and he doesn't have to spend four years in further education before earning a wage. In the 1981 recession Mark will lose his job. He will try to use his mechanical ability to open a car parts business. With little collateral and high interest rates, he will not be able to obtain a bank loan. Given the competition the car parts industry will be facing from east Asia in 1981, the expected profits of the business would have been too low for a bank to make a loan to help launch the project. He will move south to another factory. This one is non-union, and he will make less money than he did in Gary. He will stay there, despite the erratic shifts that mean he will sometimes have to work all night, and despite the increasing pain in his shoulders. In 2008, during the recession, his factory will replace him with a KUKA Robotics Corporation Titan industrial robot.

In 1975 Stephanie, both of whose parents are doctors, decides she will attend Indiana University Bloomington, majoring in psychology. Afterwards she will work for a large financial corporation in Chicago and, after a series of promotions, will become a vice president for human resources. She will invest her savings in the stock market, which will yield an average return of more than 10% for many years, and will benefit from government tax cuts that favour high earners.

These four people had very different life outcomes. Is there anything wrong with that? Each of the four made good choices knowing what they could have known at the time, everybody worked hard, and yet they had very different lives. We might say that in the card game of life they simply drew different hands.

Their parents are an important difference in the hands that they drew. This starts with the fact that Yichen and Renfu were born in China, and Mark and Stephanie in the US. The parents of the two in China were likely to be equally poor, although Communist party members enjoyed a higher level of social prestige and education. The gap in wealth between the two sets of American parents would probably have been larger. If Mark had been African American the gap would have been greater still, but his family would still have been far better off in material terms than both the Chinese families.

In 2010 the children of Stephanie and Renfu, who have been relatively successful in each country, will have access to a variety of opportunities not available to the kids of Yichen and Mark. In China, Renfu's children will attend better schools and have better job prospects because of their father's connections. With luck, they may even attend a US university, gain valuable work experience in the university-trained, English-speaking global labour market, return to China with salaries many times those of the average Chinese citizen, and be able to join the Communist Party.

Yichen's daughter will not obtain a high-quality primary or secondary education, and she may not see her mother for many years at a time. This is because the hukou restrictions mean she must go to school in Yichen's rural home district and not in the city where her mother works. Nevertheless, most likely she will be better off over her lifetime than her parents, and will certainly be better off than her grandparents.

In the US Stephanie's children will attend either a public school in her expensive neighbourhood, well funded by local property taxes, or an expensive private school. They will get early access to a much larger vocabulary, form lifelong friendships with other kids from their privileged background, and engage in a variety of interesting extracurricular experiences that help their educational performance and will help them get admission to elite universities. If they are admitted to university, the contacts, credentials and skills they will gain at the university will translate into average lifetime earnings of close to $800,000 greater than the earnings of those whose education finishes at high school level.

Mark's children will have to deal with poorly-funded public schools, the absence of union jobs, a minimum wage that will be worth less in real terms than it was in their parents' generation, and changes in technology and trade that will amplify the effects of these problems. The life trajectories of these four people illustrate just a few of the global changes in the distribution of income that have occurred in the past 40 years.

Figure 19.1 shows the distribution of income across and within countries. The height of each bar in the chart varies along two axes: the first, from left to right of the figure, is a ranking of countries from the poorest in *gross domestic income per capita* (the Democratic Republic of Congo) on the left of the figure to the richest (Norway) on the right; the second, from the front to the back of the figure, shows the distribution of income from poor to rich within each country.

The width of each country's bars represents the population of the country. The distance marked x on the top left of the figure is equivalent to a population of 200 million. For example, you can estimate that the US has a population of around 300 million, because its bar is about 1.5 times the gap between vertical lines.

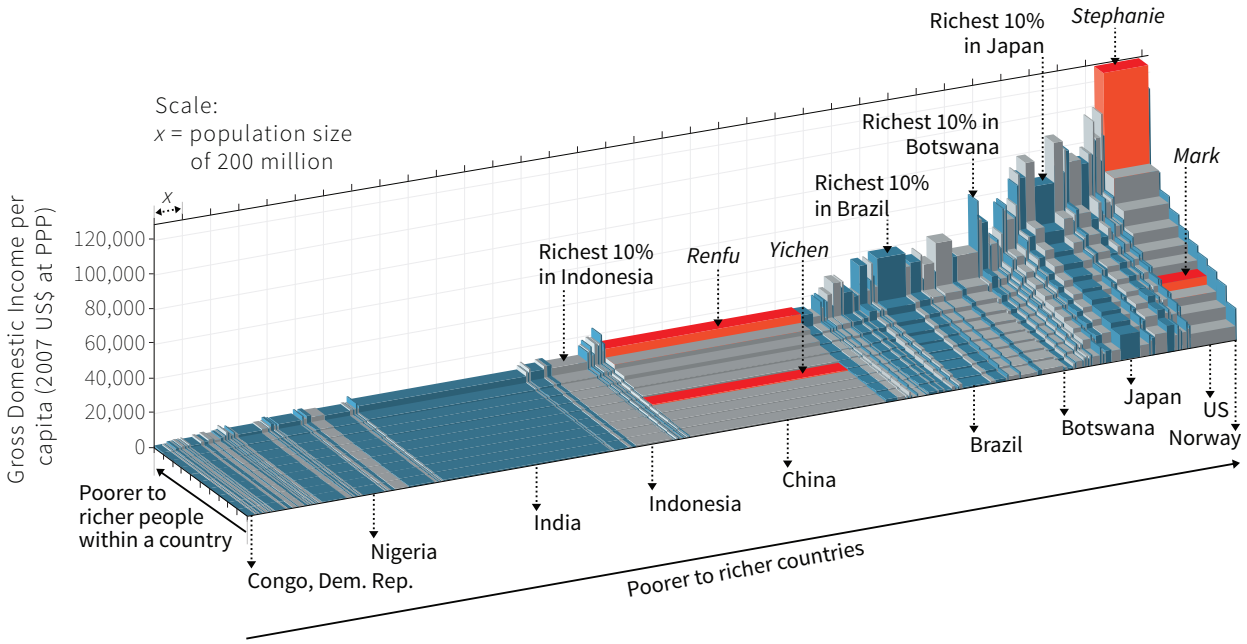


Figure 19.1 *The distribution of income in the world in 2007. Height of the bars is the gross domestic income per capita (measured in purchasing power parity dollars) of the population decile indicated.*

Source: *Bob Sutcliffe designed the representation of global inequality in this figure and supplied the data. A first version was published in: Sutcliffe, Robert B. 2001. 100 Ways of Seeing an Unequal World. London: Zed Books.*

The skyscrapers (the highest columns) at the back of the right-hand side of the figure represent the income of the richest 10% in the richest countries. The tallest skyscraper is the richest 10% in the US. This exclusive group has gross domestic income per capita of a little more than $125,000. Although Norway has the highest gross domestic income per capita and is therefore the country at the right-hand end of the figure, it does not have a particularly tall skyscraper for the richest 10% (our view is almost entirely blocked by the US skyscraper) because income is more evenly distributed in Norway than in some other rich countries.

We have marked the positions of Renfu, Yichen, Stephanie and Mark in the world income distribution on the Figure 19.1.

Their stories also illustrate that inequality in economic dimensions such as wealth and income are often associated with disparities in education, in health, and in the scope of choices open to a person. While both Stephanie and Mark face trade-offs in pursuing their life objectives, Stephanie's feasible set is a larger than Mark's. The size of her feasible set is an indicator that Stephanie is better off than Mark. For example, she could have both more income and more free time than Mark if she decided that's what she wanted to do. Because she has more choices, she is freer than Mark.

While inequalities in freedom are an important subject, in this unit we will focus on inequalities in wealth and income.

19.1 TECHNOLOGY, ENDOWMENTS, INSTITUTIONS AND INEQUALITY

The inequalities among the four people whose stories we told are partly the result of the choices they made, and their individual capacities and objectives. But many of the differences are not really about them as individuals, but instead about where, when and to whom they were born. Stephanie, for example, might have done well had she been born in China, but she almost certainly would have ended up earning less than Mark did in the US until he was laid off. In short, inequalities depend not only on the people, but also on their circumstances.

The factors influencing the extent of economic inequality can be understood using a simplified model representing the cause-and-effect relationships (which we show by arrows from cause to effect) in Figure 19.2.

The figure shows that both technology and institutions influence economic inequality directly, but also indirectly through their effect on endowments. Technology and institutions also affect each other—as we saw for example in Units 1 and 2, where the competition for innovation rents that is a part of the institutions of a capitalist economy accelerated the rate of adoption of new technologies. We will return to this model and provide examples of each of the arrows as a summary at the end of the unit.

Changes in inequality can feed back to affect technology, institutions, and endowments (though we do not show these arrows). As the lives of our four hypothetical characters illustrated, an increase in economic inequality in one generation affects the endowments of the next and, in turn, inequality in the generation to follow.

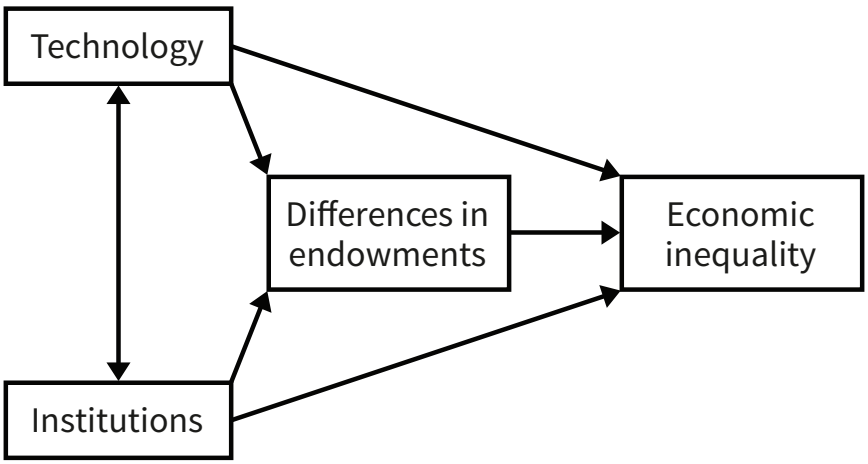


Figure 19.2 *The causal relationships among technology, institutions, endowments and inequality.*

Endowments

The first set of factors determining the extent of economic inequality among people is the differences in what they have or what they are—differences in their wealth and in their skills and other capacities. Facts about an individual that may affect his or her income are called *endowments*. Endowments are the basis of wealth.

ENDOWMENT

Facts about an individual that may affect his or her income:

- The physical wealth a person has is a component of that person's endowment
- A person's CV is a partial list of the individual's endowment, because it shows schooling, special training, work experience in internships, the person's citizenship and whether the individual has a visa (or green card) allowing employment in a particular labour market
- When race or social class background affect the probability of getting a job, it means they are endowments for that applicant
- The nonmaterial aspects of the broader definition of wealth, such as schooling and training, are sometimes called *human capital*

DISCUSS 19.1: TECHNOLOGY, INSTITUTIONS, ENDOWMENTS AND INEQUALITY

Choose a particular country and time period that you know well. For your chosen country and time, replicate Figure 19.2 above, with specific examples of each of the boxes and arrows, to show the effect of technology and institutions on differences in endowments (be as specific as possible about examples of endowments) and inequality. You can use the table in Figure 19.20 to help you.

Employers scour CVs of job applicants for evidence of endowments: nationality, gender, even the person's race or social class. Social networks now allow employers to research other endowments, including a person's appearance and connections.

To see how inequalities may arise because of who one interacts with, and on what terms, think back to the interactions you have studied in the previous units, summarised in Figure 19.3. In the table we present the cases in the first column, starting with the interaction in Unit 5 between Bruno, the landlord, and Angela, the farmer.

Recall that how much Bruno got depended on:

- *The fact that Bruno owned the land*: He could exclude Angela from working it (Bruno's endowment).
- *Angela's productivity as a worker*: This is determined by Angela's endowment and the available technology.
- *Angela's reservation option*: What Angela would get if she were to refuse to work for Bruno or he refused to hire her. It is determined by her endowments and the institutions or policies in place.

Endowments and classes

The endowments of the pairs of individuals in Figure 19.3 appear in the second column of the figure, starting with Bruno owning the land and Angela owning only her time and capacity to work. This inequality in land ownership matters because it determines who has to work for whom, and who can earn income from allowing others to work with their capital good or their land.

Endowments matter in another way, because they change Angela's reservation option (in this case, what she would get if her employment relationship with Bruno ended). If Angela owned land that she could work herself, then Bruno would need to pay her at least enough that she would rather work for him than work on her own land.

SITUATION, ACTORS AND UNIT	ENDOWMENT	RESERVATION OPTION	CONFLICT OVER?	INSTITUTIONS AND POLICIES (EXAMPLES)	TECHNOLOGY (EXAMPLES)
LANDLORD AND FARMER: BRUNO AND ANGELA *UNIT 5*	Bruno owns the land Angela has 24 hours of potential labour	Bruno: rent to another farmer Angela: government support	Rent paid by Angela to Bruno and the hours that Angela works	Angela's reservation option and legislation limiting work hours	Angela's increased productivity due to an improvement in seeds allows Bruno a larger surplus when he has all the bargaining power
BORROWING, LENDING AND INVESTMENT: JULIA AND MARCO *UNIT 11*	Julia: $100 next year Marco $100 now	Julia: consume nothing now, $100 later Marco: consume some now, store and consume some later	Julia benefits from a low interest rate and Marco benefits from a high interest rate	Competition among lenders and interest rate regulation	An improvement in storage technology makes it easier for Marco to move his goods forward in time, and also raises the rate of return on his investments
SPECIALISATION AND TRADE: GRETA AND CARLOS *UNIT 16*	The skills and resources of each that determine their feasible consumption set in the absence of specialisation and trade	Both: the utility they would enjoy if they did the best possible without trading	Price at which they exchange the good in which they specialise	Price-setting power by either Greta or Carlos	An improvement in the technology of the good in which one specialises will benefit both countries, the larger gains going to the person with price-setting power
FIRM: OWNER AND EMPLOYEES *UNIT 6*	Owner: ownership of the firm Employee: capacity to work given her skills	Owner: hire some other employee Employee: unemployment insurance and job search	Wage, working conditions, effort on the job	Level of unemployment insurance, employment level and legislation regulating work conditions	A new technology may increase the productivity of the employee's effort, transferring a greater surplus to the employer (short run) and increasing employmentand the real wage (long run). It may also affect how easily the employer can monitor the employee's effort
BANANA PLANTATION: OWNERS AND DOWNSTREAM FISHING COMMUNITIES *UNIT 10*	Owners: the land and other capital goods of the plantation. Fishing communities: their boats and capacity to catch fish	Owners: raise bananas without using Weevokil pesticide. Fishers: convert to farming	Use of polluting chemical, possible compensation for destruction of fisheries or commitment not to use Weevokil	Regulations governing the use of pollutants and enforcement of private agreements made between the parties	A new pesticide technology could reduce or increase the conflict between the two groups depending on its external effects

Figure 19.3 *Inequality: Endowments, reservation options, conflicts, institutions and technologies.*

Differences in endowments also explain why Julia and Marco are on the opposite sides of the credit market. Marco has a sum of money now and Julia has none, so Julia will borrow from Marco and repay him with interest. Similarly, the owner and employees of a firm are on different sides of the labour market. If the workers owned the buildings, machinery and other assets making up the firm, they would most likely not be employees. We say that Julia and Marco, or the owner and employees are in different *classes*.

CLASSES

Groups of people who, because of their differing endowments, engage in asymmetric economic interactions with members of other groups, such as:

- Owners and employees
- Landlords and tenants
- Borrowers and lenders

Economic inequality exists both between classes (as in these examples) and also within classes, as when employees with distinct skills or differing by ethnic group are paid different wages or for whatever reason are out of work.

We call these interactions *asymmetric* because the actions open to one party—the employer for example—are not open to the other—the employee. The employer, for example, sets the wage and the tasks that a worker is directed to perform, and can also fire the worker. The worker chooses how to go about her work within limits that the employer sets. These interactions differ from those we studied in Unit 4, for example, where the actions open to all parties were identical—for example, use integrated pest management or chemical fertilizers, learn C++ or Java.

Asymmetric relationships among classes are based on the different endowments that the members of different classes possess, and are associated with differences in income, but also in power. The employer can fire the worker and deprive her of the employment rent she would otherwise receive. The fear of losing this rent is an important reason for the worker to carry out the employer's wishes.

(The worker could, of course, quit. But this does not make the relationship symmetrical. If she is receiving a rent she would penalise herself by quitting, and her employer would just replace her with someone currently unemployed.)

Interactions between members of different classes are not only economic—involving mutual gains through exchange and conflicts over their distribution—they are also political—involving the exercise of power by one party over the other.

DISCUSS 19.2: CLASSES, ASYMMETRY AND CONFLICT

Consider the first (landlord and farmer) and last (banana plantation owner and downstream fishing communities) relationships listed in Figure 19.3.

1. Identify the classes on either side of the relationship and the asymmetries in their economic interactions.
2. For each of the institution/technology examples listed in the last two columns, explain how the relationship would be affected.

Yichen, Renfu, Stephanie and Mark illustrate these differences in power: At the top of large corporations, both Renfu and Stephanie regularly give orders to others with the expectation that they will be obeyed. When Yichen works at the Motorola plant, and when Mark worked at the factory until he was replaced by a robot, they would have little choice but to carry out the directives of the owners and managers who employed them.

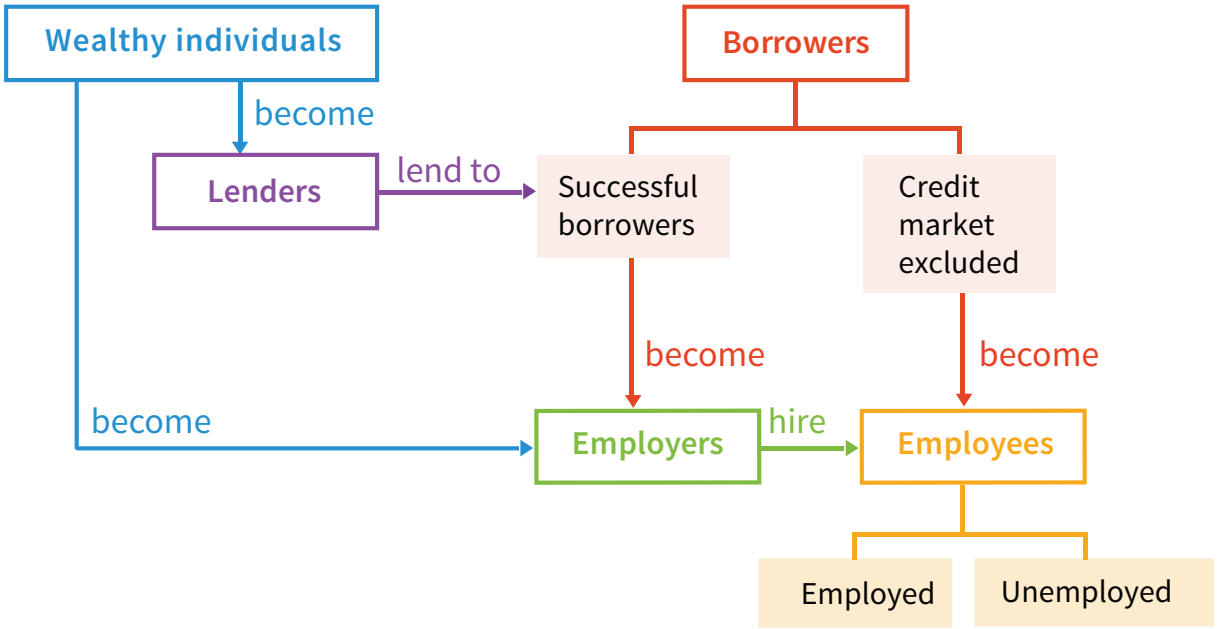


Figure 19.4 *The credit and labour markets shape the relationships between classes.*

Figure 19.4 illustrates how the credit and labour markets together influence the relationships among the classes of lenders and borrowers and employers and employees. The credit market allows those without wealth to smooth consumption (as we have described in Unit 11 and Unit 12), and also to acquire money to purchase capital goods to become an employer. In this model, those who are wealthy enough

to have substantial endowments in the form of capital goods (either by owning them outright or by borrowing) become employers. Those who are not wealthy become employees or unemployed.

Institutions and technology

The value of a particular endowment, say a programming skill or ownership of a 3D printer, depends on both *technology* and *institutions*. Being physically strong was a valuable endowment in agriculture—at least until mechanisation made it less important in determining earnings. In that case a change in technology reduced the demand for a particular kind of skill, and so its value (relative to other skills) fell. The wage that a farm worker received may have been decided in three ways:

- It could have been decided by supply and demand for the worker's endowments.
- It might have been set by law.
- It might have been decided by the bargaining power of a trade union.

Therefore, institutions matter in this case.

We saw another example of how institutions affect the value of a person's endowment in Unit 11 and Figure 19.3. Recall that Julia's endowment is $100 next year. Her wealth now (what would be measured in the bathtub) depends on the institutions determining whether she can borrow, and the interest rate at which she can borrow. If her only option is the village moneylender in Chambar, she faces a high interest rate and her wealth (now) is much lower than $100; if she can borrow at a low interest rate, her wealth is quite close to $100. If she can't borrow at all, then there is nothing in the bathtub: her wealth now is zero.

If your endowment is a scarce machine that you own, which produces something that many people (and even better, many rich people) want, then many firms will want to rent that machine from you, and so it will command a high price. In the last column of Figure 19.3 we consider the role of changes in technology in affecting the degree of inequality. In the row concerning the firm's owners and employees, a labour-saving technology, as we saw in Unit 15, can—at least initially—reduce the number of workers a firm needs, making employees more vulnerable to job loss and reducing the likelihood of getting another job at the same wage for those who have been fired.

Recall that a change in institutions can change Angela's reservation option. Before the rule of law, the institutions were such that Bruno could simply coerce her to work, and the only thing that constrained the size of the surplus he could get was the need to keep Angela healthy enough to work the next day.

The institutional change, which gave her the right to say no, changed this reservation option: Bruno had to offer Angela a deal that would make her better off working for him than not working. Institutions can also alter the set of allocations that were allowed (for example, by restricting the maximum length of Angela's working day).

While this outcome was Pareto-inefficient, as we saw in Unit 5 (Figure 5.10), it also reduced inequality. We will examine a variety of policies at the end of this unit, and discuss how they influence equality and efficiency.

DISCUSS 19.3: YICHEN, RENFU, MARK AND STEPHANIE

Consider the economic situation of Yichen, Renfu, Mark and Stephanie, discussed at the start of this unit. Give examples of how technology, institutions and differences in endowments explain any differences in economic success between these actors.

19.2 INEQUALITY BETWEEN EMPLOYERS, THE EMPLOYED AND THE UNEMPLOYED

To study inequality between employers and employees we return to the model of the labour market introduced in Units 6, 9 and 15. Remember there are two classes—employers and employees—and, among the second class, some are typically unemployed.

In Unit 6 we used the wage curve to explain the relationship between employment (the fraction of the labour force that is employed) and the wage rate offered by firms, and how this relationship is affected by changes in the employees' reservation option (for example, changes in the unemployment benefit).

In Units 9 and 15 we showed how the wage curve and the profit curve determine the level of employment and the wage rate in the economy taken as a whole.

As we did in Unit 1, we can construct the *Lorenz curve* and calculate the *Gini coefficient* for the economy in this model. Read the Einstein sections called *Inequality as differences among people* and *The Lorenz curve, the Gini coefficient and a small population bias* that explain how to calculate the Gini coefficient with different kinds of information about a population.

In Figure 19.5a we see an economy with 80 employees of 10 firms. Each firm has a single owner. There are also 10 unemployed people. Because the unemployed people receive no income (we will introduce an unemployment benefit shortly), the Lorenz curve (the solid blue line) begins on the horizontal axis to the right of the left-hand corner. In total, the 80 employees receive 60% of the income as their wages, while

the firm owners receive the remaining 40%. The size of the shaded area measures the extent of inequality, and the Gini coefficient is 0.36. To learn how to calculate the Gini coefficient from information like this on the composition of the population and the wage rate, see the Einstein section called *The Lorenz curve and the Gini coefficient in a class-divided economy with a large population*.

Figure 19.5a *The Lorenz curve among employers, employed and the unemployed.*

The Lorenz curve is made up of three line segments with the beginning point having coordinates of *(0, 0)* and the endpoint *(1, 1)*. The first kink in the curve occurs when we have counted all the unemployed people, so everyone else has some income. The second is the interior point, whose coordinates are *(fraction of total number of employees, fraction of total output received in wages)*. The fraction of output received in wages, sometimes called the wage share in total income, s, is

$$s = \frac{\textit{wage per hour}}{\textit{output per hour of labour}}$$

Therefore the shaded area in the figure—and hence the Gini coefficient—will increase if:

- A larger fraction of the employees is without work
- Wages fall and nothing else changes
- Productivity rises and nothing else changes (wages do not rise)

We now study how changes in the three factors influencing inequality of income in Figure 19.3—institutions, endowments and technology—result in changes in the distribution of income represented by the Lorenz curve and the degree of inequality as measured by the Gini coefficient.

Institutions

In Unit 5, you studied how a change in institutions—a reduction in the maximum allowable rent charged by landlords in the Indian state of West Bengal—can affect the Lorenz curve and the Gini coefficient. Differences in institutions governing the relationship between employers and workers will be associated with differences in the degree of inequality.

Recall that, among the incentives that firms provide for workers to work hard and well, is the payment of a wage higher than the worker would be able to get if she was fired, and the threat that inadequate effort or care on the job may result in the workers becoming unemployed. But trade unions often resist what are called *for cause* firings, and national legislation often makes dismissing employees a costly process for firms.

Figure 19.5b depicts the effects of a change in legislation that has made it easier for a firm to fire a worker who is not conforming to management's requirements for the pace of work. The greater likelihood that a shirking worker would be fired under the new legislation makes it cheaper for the employer to get his employees to work, as we saw in Unit 6. If output remains unchanged (no new firms enter), the employers now find they can get workers to work for a lower wage. The combination of a lower wage for the workers and higher incomes for the owners means the Gini coefficient rises from 0.36 to 0.47.

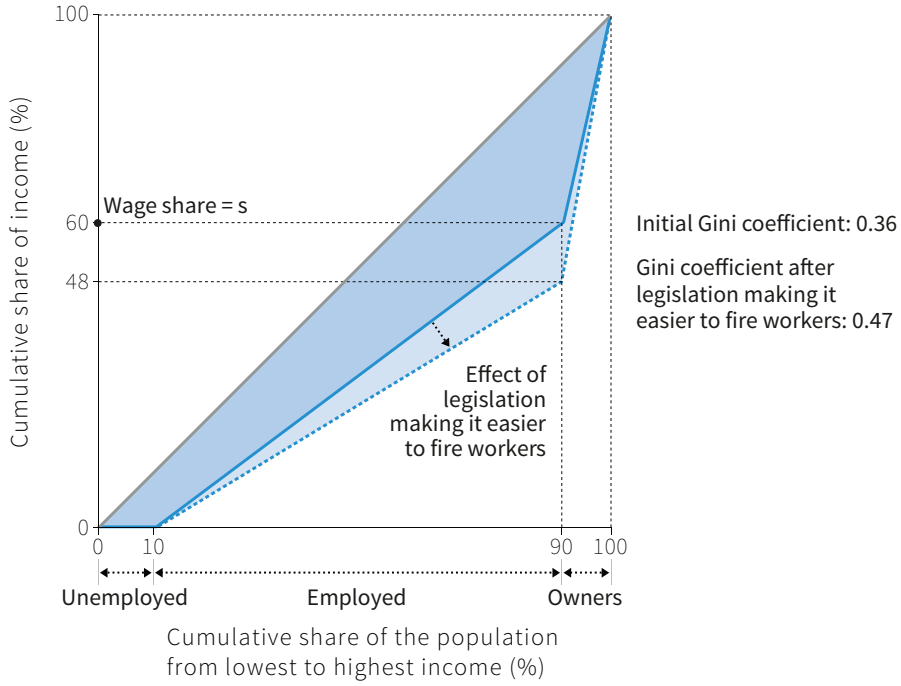


Figure 19.5b *The effect of making it easier to fire workers on inequality among employers, employees and the unemployed (immediate effect, assuming output constant).*

In the model introduced in Unit 15, however, this effect will not be permanent because a higher rate of profit will motivate firms to hire more, lowering the unemployment rate and raising wages (due to the upward-sloping wage curve), and reducing inequality.

DISCUSS 19.4: UNEMPLOYMENT INSURANCE

Start with the initial situation in Figure 19.5a. Now suppose that following a change in policy, the unemployed receive *unemployment insurance* equal to 50% of the wage.

1. Assuming that output and unemployment remain unchanged, and that the Gini coefficient before the change in policy was 0.36 as in Figure 19.5a, calculate the Gini following the introduction of unemployment insurance.
2. Draw the Lorenz curves before and after the policy was enacted.
3. Provide an explanation of why the introduction of unemployment insurance reduces inequality.
4. Suggest why the assumption that output remains unchanged may not hold.

Endowments

The employee's primary endowment is a capacity to work rather than physical wealth, so now consider a case in which workers have acquired more schooling, which increases productivity (a unit of effort now produces more goods). This means that, at the previous wage, the greater productivity of workers now produces a higher profit for the firms.

Figure 19.5c shows what happens if, with higher profits, firms enter. Production expands, which reduces the rate of unemployment. Because this increases the workers' reservation position, it raises the wage.

Note that because total output has increased (more workers are employed and they are more productive), the fact that owners of the firms now receive a smaller share of the total output is consistent with their profits having risen.

Figure 19.5c *The effect of a more educated workforce on inequality among employers, employees and the unemployed.*

Technology

To see how changes in technology affect the degree of inequality, consider what happens if a labour-saving technology is introduced and 10 of the employed people are no longer needed in production, so they join the unemployed. As a result of the reduction in employment (the increase in unemployment) the reservation position of the remaining workers has deteriorated: if they were to lose their job they would remain unemployed for longer. Knowing this, the firms can now pay a lower wage, so the remaining 70 employees now receive only 42% of the output, the firms gaining the rest. The lower wage of the employed workers is indicated by the flatter dashed Lorenz curve for the portion relating to the employed workers. The increased income of the owners is reflected in the steeper line indicating their income.

The resulting Gini coefficient has risen from 0.36 to 0.56. The effect will not be permanent, however; as we saw in Unit 15, improvements in technology result in job creation as well as job destruction. In that model, firms will enter, production will expand and some of the unemployed will get jobs.

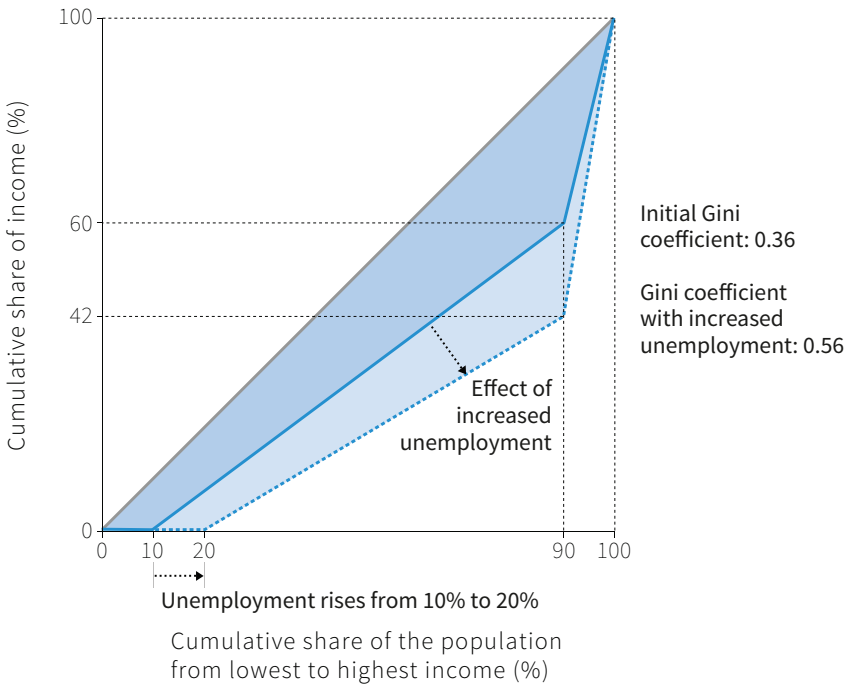


Figure 19.5d *The immediate effect of a labour saving technological change on inequality (no increase in output).*

19.3 INEQUALITY AMONG LENDERS, BORROWERS, AND THOSE EXCLUDED FROM CREDIT MARKETS

Long before there were employers, employees and the unemployed, there were lenders and borrowers. The class of borrowers includes those excluded entirely from the credit market. Differences in income between those who lend and those who borrow—people like Marco and people like Julia—remain an important source of economic inequality today.

We can analyse inequalities between these classes (and among the borrowing class) using the same Lorenz curve and Gini coefficient model.

Here is an illustration. An economy is composed of 90 farmers who borrow from 10 lenders and use the funds to finance the planting and tending of their crops. The harvest (on average) is sold for an amount greater than the farmer's loan, so that for every euro borrowed the farmer gains income of $1+\rho$, where ρ is called the *rate of profit*.

Following the harvest, the farmers repay the loans with interest, at rate i. We simplify by assuming that all of the loans are repaid and that all lenders lend the same amount to the farmers at the same interest rate.

Because for each euro borrowed, total revenue is $1+\rho$ and the cost of inputs is 1, it follows that income (revenue less costs) in the total income of the economy is proportional to ρ. Income is divided between the lender who receives an income of i for every euro lent, and the borrower who receives the remainder, namely i-ρ. The share of the total output received by the lender is i/ρ and by the borrower is $1 - i/\rho$.

Thus, if we have $i = 0.10$ and $\rho = 0.15$ then the lender's share of total income is 2/3 and the borrower's is 1/3.

Inequality in this economy is depicted Figure 19.5e. The Gini coefficient is 0.57.

Recall that in Unit 11 we showed why some would-be borrowers—those unable to post collateral or lacking their own funds to finance a project—may be excluded entirely from borrowing even if they would be willing to pay the interest rate. How does this affect the Lorenz curve and the Gini coefficient?

To explore this, imagine that 40% of the prospective borrowers are excluded (and since they cannot borrow, they receive no income at all) and that nothing else in the situation changes (i and ρ remain unchanged).

The new situation is shown by the dashed line in Figure 19.5e. The new Gini coefficient is 0.70, showing an increase in inequality as the result of the credit market exclusion of the poor.

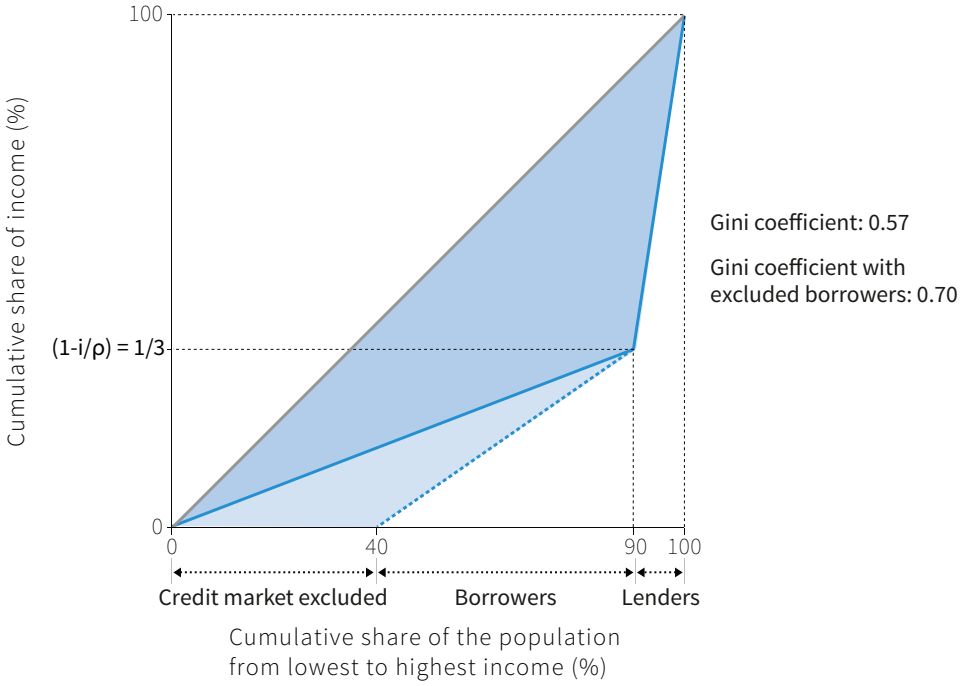


Figure 19.5e *Inequality in a borrowing and lending economy. Note: The Gini coefficient when there are no borrowers excluded is 0.57; when 40 are excluded it is 0.70.*

19.4 INEQUALITY ACROSS THE WORLD AND OVER TIME

Economists and statisticians use Lorenz curves to estimate Gini coefficients so that they can measure inequality in wealth, income, earnings (income from work in the form of salaries and wages), years of schooling, and other indicators of economic or social success.

Figure 19.6 presents data on three dimensions of inequality—wealth, wages, and disposable income—in three economies. Recall that disposable income is the income that a family can spend after paying taxes and including any monetary transfers from the government such as unemployment insurance and old age pensions. Two things stand out:

- *Wealth is much more unequally distributed than earnings, which in turn are much more unequally distributed than disposable income*: Though the differences among the three are much smaller in Japan than in Sweden and the US.
- *The remarkable equality of Sweden's disposable income*: This is due to its modest inequality in earnings and system of taxes and transfers, not to relative equality in its distribution of wealth, which is similar to that of the US.

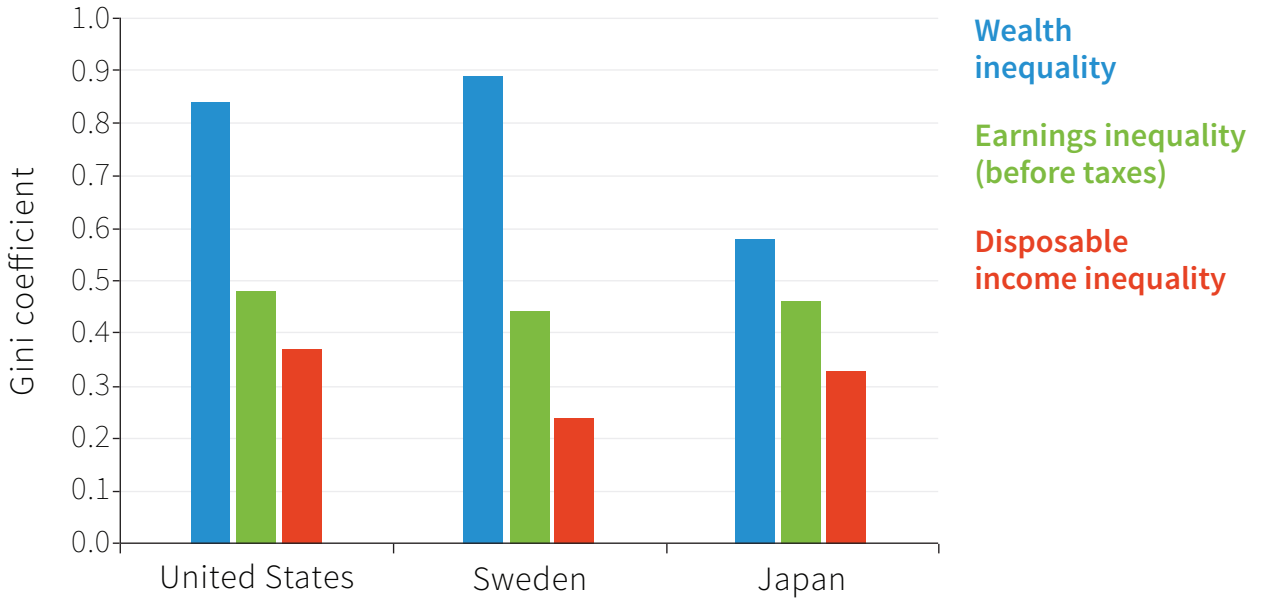


Figure 19.6 *Inequality in wealth, earnings and disposable income: US, Sweden and Japan.*

Source: *Fochesato, Mattia, and Samuel Bowles. 2013. 'Wealth Inequality from Prehistory to the Present: Data, Sources and Methods.' Dynamics of Wealth Inequality Project, Behavioral Sciences Program, Santa Fe Institute; Fochesato, Mattia, and Samuel Bowles. 2013. Technology, Institutions and Wealth Inequality in the Very Long Run. Santa Fe Institute; Wang, Chen, and Koen Caminada. 2011. 'Leiden Budget Incidence Fiscal Redistribution Dataset.' Version 1. Leiden Department of Economics Research.*

DISCUSS 19.5: INEQUALITIES AMONG YOUR CLASSMATES

1. Using this Gini coefficient calculator, calculate the degree of inequality of height among your classmates.
2. Why is this Gini so much smaller than it was for wealth in the data above?
3. Now use the calculator to compute the Gini for another measure (for example, age, weight, commuting time to university, number of siblings or grade in the last exam).
4. Explain any differences between this Gini and that for wealth.

Another way to measure inequality focuses on the very rich, providing an answer to the question: what fraction of total income or wealth is accounted for by the richest 1% or 10% of the population? This indicator has the advantage that it can be measured over hundreds of years, because the very rich have long been required to pay taxes, and hence we have reasonably good information on their incomes and wealth. Figure 19.7 shows the fraction of all wealth held by the richest 1%, for all countries on which data is available.

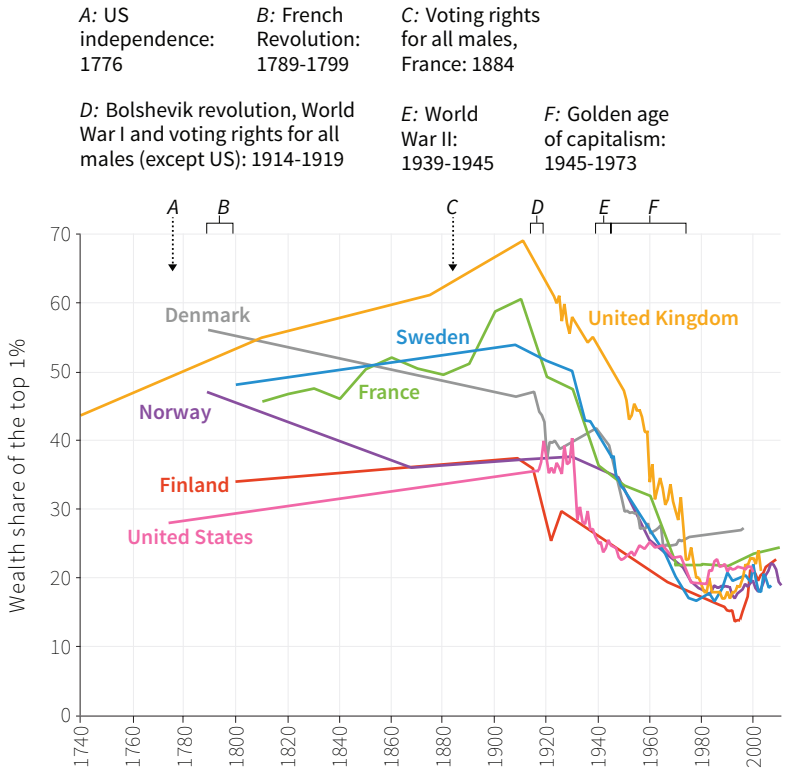


Figure 19.7 *Share of total wealth held by the richest 1%: 1740-2011.*

Source: *Adapted from Figure 19 of Waldenström, Daniel, and Jesper Roine. 2014. 'Long Run Trends in the Distribution of Income and Wealth.' In Handbook of Income Distribution: Volume 2a, edited by Anthony Atkinson and Francois Bourguignon. Amsterdam: North-Holland. Data.*

There appear to be three distinct periods: the 18th and 19th centuries up to about 1910 show increasing wealth inequality (excepting Norway and Denmark), the 20th century until 1980 shows decreasing wealth inequality, and the period since shows a modest increase in wealth inequality.

Figure 19.8 presents similar data but for the share of income before taxes and transfers (rather than wealth) received by the top 1% of income earners. As in Figure 19.7 there are country differences: in recent years the US is much more unequal than China, India, or the UK for example. But there are also common trends, similar to the second and third periods in the distribution of wealth: a trend towards less inequality in much of the first three quarters of the 20th century followed by an increase in inequality since about 1980. We will see later that this U-turn did not occur in all countries, including most of the major economies of the continent of Europe.

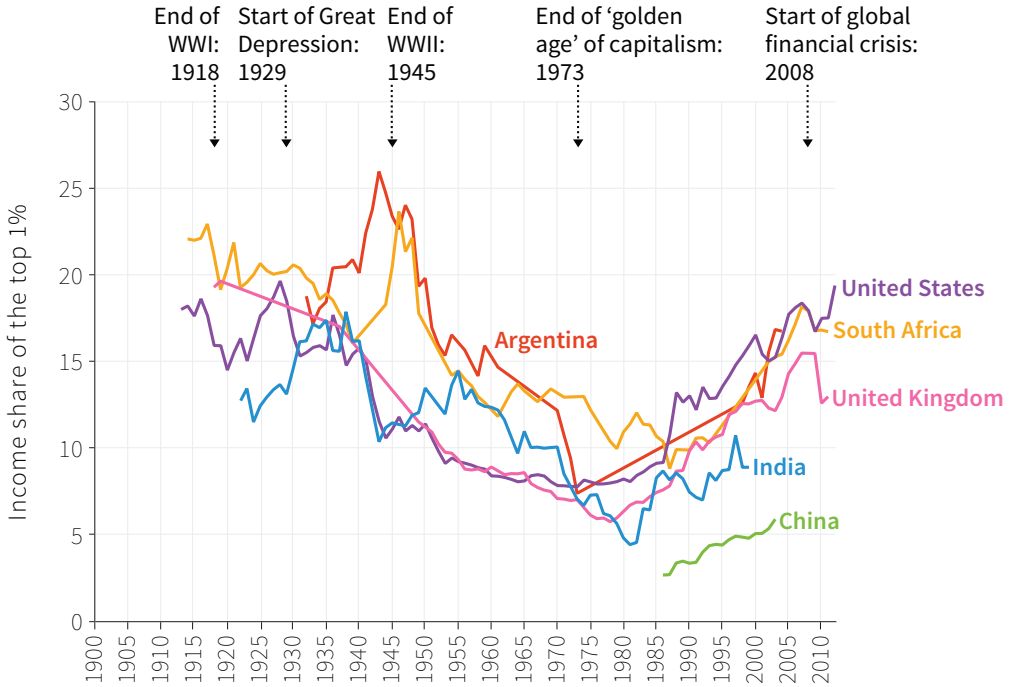


Figure 19.8 *The share of total income received by the top 1%: 1913-2012 (for an interesting comparison look ahead to Figure 19.18).*

Source: *Alvaredo, Facundo, Anthony B Atkinson, Thomas Piketty, Emmanuel Saez, and Gabriel Zucman. 2016. 'The World Wealth and Income Database (WID).'*

We used data created by Thomas Piketty and his collaborators to create Figures 19.8 and 19.13. He is an economist and author of the bestselling economics book Capital in the 21st century. In this Economist in Action video, he examines economic inequality from the French Revolution to the present, and explains why careful study of the facts is essential.

19.5 GROUP INEQUALITY

From Figure 19.1 it appears that that the most important determinant of income is the country in which you were born. If you doubt this, try the following thought experiment. All you care about is income. Do you choose option 1 or option 2?

1. You can choose the decile you are to be in, but not the country.
2. You can choose the country you are born and live in, but not the decile.

If you chose option 1, you would of course choose to be in the top decile, so you would be at the back of Figure 19.1. But you have an equal chance of being on the left-hand side or the right-hand side. If you chose option 2, you could select one of countries at the right-hand end with the highest average income. You are as likely to be in the lowest decile, at the front of the figure, as in the highest decile.

Figure 19.9 shows the income inequality among the world's population irrespective of the country they live in (the blue dots) as measured by the Gini coefficient (income is measured before taxes and transfers). Also shown is the amount of income inequality that would result if, hypothetically, everyone in each country had the same income—the average for that country. If this were the case, then the only source of inequality in the world would be inequality between countries.

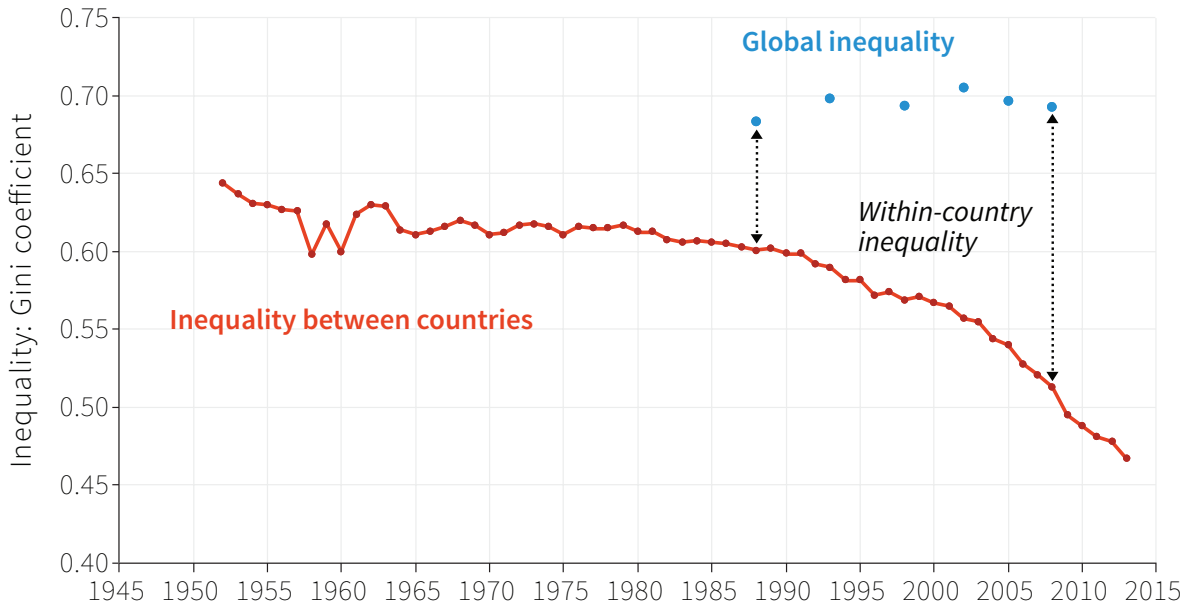
You can see in 1986 the Gini coefficient for all families in the world (the first blue dot) was 0.68, and this number would have been 0.60 had there been absolutely no inequality within each country (the red line). As a result, we see that 88% of global inequality in income is accounted for by our measure of inequality between countries (that is because $0.60/0.68 = 0.88$, or 88%).

The figure also shows that between-country inequality has been falling rapidly: by 2008, 74% of global inequality was between-country inequality ($0.51/0.69 = 0.74$).

At the beginning of Unit 1 we wrote that, prior to the emergence of capitalism:

"The prospects of a daughter or a son depended on where their parents were on the economic ladder. It mattered much less in which part of the world the son or daughter was born."

The economic take-off of the first capitalist economies changed this for North Americans and Europeans, as even the poor in these countries became richer than even the rich elsewhere. This is now changing again, as we have seen, as a result of the economic take-off of India and China.



Global inequality (1986 to 2008)

The blue dots show income inequality among all families in the world. It is, effectively, the world's Gini coefficient.

The hypothetical inequality between countries falls...

The red curve shows the between-country income inequality between 1952 and 2013. To calculate it we assume everyone in a given country had the same income. It started to decline rapidly the 1980s.

... and within-country inequality rises

The decline in between-country inequality accelerated as the growth of the world's largest poor countries, China and India, increased. In 1986, 88% of global inequality was between countries. By 2008, this was 74%.

Figure 19.9 *Group inequality: Global and between-country income inequality (1952-2013).*

Source: *Milanovic, Branko. 2012. 'Global Income Inequality by the Numbers: In History and Now -an Overview-.' Policy Research Working Papers. The World Bank.*

It is still the case, however, that the greater source of inequality in the world today is the *group inequality* based on country of citizenship. Passports and borders limit the economic opportunities people from different countries face. People with the same level of education, intelligence and ambition, but born on different sides of a national border, face very different life chances, whether that is the border between Mexico and the US, North and South Korea, or the Mediterranean Sea that divides North Africa from Europe. Even where migration is allowed, migrants are often denied access to political and labour rights, as occurs in the Gulf and some East Asian countries.

Group inequality also exists within countries. Vast disparities in life chances in India, for example, follow from 2,000-year-old hereditary caste boundaries. South African apartheid formalised inequality with a complex system of racial barriers. Many countries, for example Australia, the US and much of Latin America, have high inequality between descendants of colonists and indigenous people.

Some countries have been endorsing equal rights for hundreds of years, but have not eliminated economic inequalities between men and women. Figure 19.10a shows the expected lifetime earnings (labour income) of men and women assuming that they work full time from the time they leave school until retirement. As a result, any differences in Figure 19.10a are not due to women on average having more time out of the labour force because of child-rearing. Because the quality of schooling does not differ between males and females on average (and girls tend to do as well on most tests) the gender differences in pay are not due to differences in cognitive ability, or quality of schooling. Yet for every level of schooling women can expect to earn less than men.

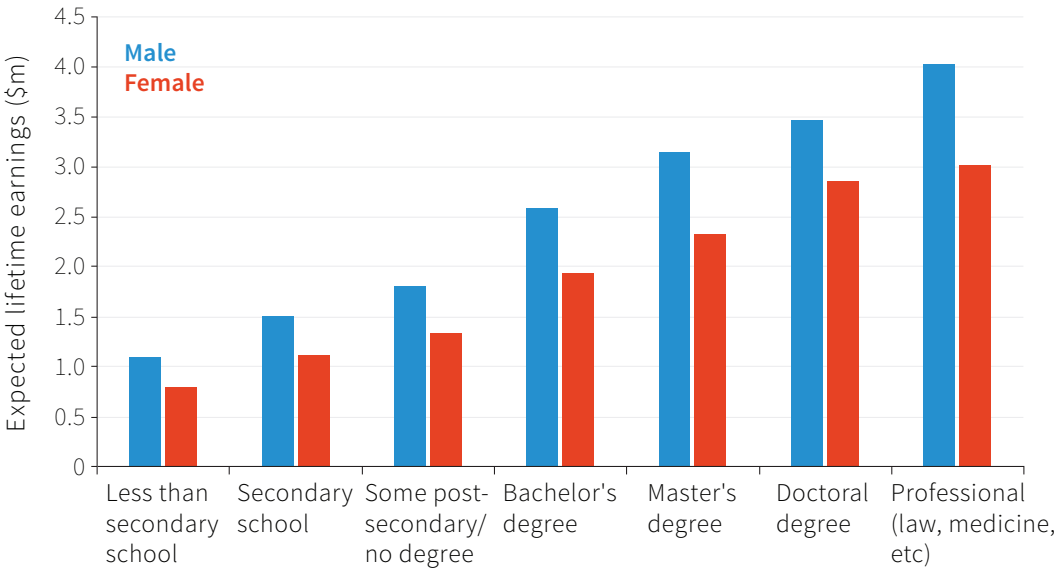


Figure 19.10a *Group inequality: Schooling and lifetime earnings for men and women in the US.*

Source: *Adapted from Figure 5 in Carnevale, Anthony P, Stephen J Rose, and Ban Cheah. 2011. The College Payoff. Georgetown University Center on Education and the Workforce.*

The figure also shows, however, that additional schooling contributes to higher lifetime incomes, and that those women who complete university (a bachelor's degree) earn much more than men who ended their schooling after secondary school. But for most of the people of the world, investing in four years of education after secondary school will have a smaller payoff than investing in a new passport by gaining citizenship of a higher-income country.

In many parts of the world girls receive considerably less schooling than boys but, as Figure 19.10b shows, girls go to school for the same time as boys in both the US and France, and longer in Brazil. China and Indonesia have virtually eliminated the gender gap in years of schooling, and India is rapidly closing it.

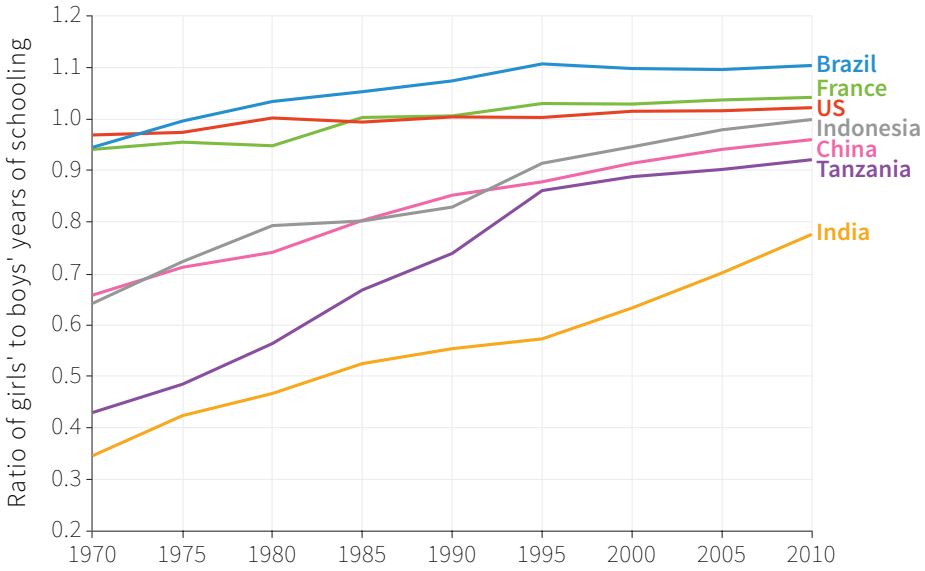


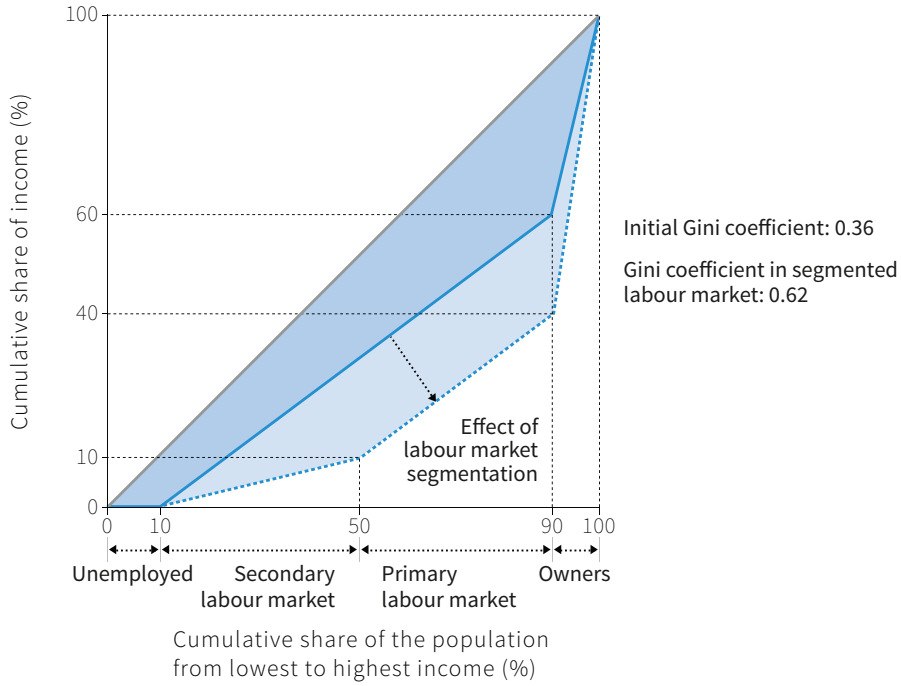
Figure 19.10b *Group inequality: Average years of schooling, girls relative to boys (1970-2010).*

Source: *The World Bank. 2016. 'IIASA/VID Educational Attainment Model. Dataset Produced by the International Institute for Applied Systems Analysis (IIASA) in Laxenburg, Austria and the Vienna Institute of Demography, Austrian Academy of Sciences.' Educational Attainment Statistics.*

To understand why some societies end up highly segregated by race or some other group characteristic, take two minutes to play the online game *The Parable of the Polygons*.

A final example of the way that institutions can influence inequality is what is termed the *segmented labour market*. Until now we have assumed that all workers receive the same wage in a single labour market, but in reality there are many distinct labour markets. In the primary labour market workers are typically represented by trade unions, and enjoy high wages and job security. Workers in the secondary labour market might be primarily young, or discriminated against by race or ethnic group, or simply workers on short-term contracts with limited wages and job security.

Figure 19.11 shows the effect of labour market segmentation, with a low-wage segment and an equal number of high-wage primary segment workers. The owners are not segmented because they can easily invest their wealth in firms in either or both sectors and, as a consequence, the rate of return will be the same in both sectors.



Labour market segmentation

This divides employees equally into low-wage and high-wage groups. In this new setting, the Gini coefficient is higher than before.

Figure 19.11 The effect of labour market segmentation (or a model of inequality in the world economy).

We can also use Figure 19.11 as a simplified way to view the world economy. Instead of two labour market segments in the same country, there are two countries, a poor one with low wages and a high-wage country, a little like China and Germany in Unit 16.

Just as it is not easy for workers to move up from the secondary to the primary labour market within a country, the global economy has nationally segmented labour markets because of the barriers facing workers who would like to relocate from one country to the other. And, just as in the national economy, owners are not segmented. They invest their wealth wherever it will get the highest return. (International capital mobility is high: money does not need a green card or a work visa to be allowed to "work" in a country.)

19.6 INHERITED INEQUALITY

In addition to differences in group membership such as your nation, sex, race, or ethnic group, a second source of economic inequality within a nation is inherited: you are rich or poor because your parents were rich or poor. In most countries 200 years ago it was taken for granted that somebody should expect a life of poverty simply because her parents had been poor, or that someone else should inherit the ownership of his father's company and social status, without having to prove that he was the best person for the job. But this has changed with the spread of public education and, in many countries, with the decline in discrimination against poor people due to their race, religion, or simply their origins. The economic status of one's parents is now not automatically transferred to the next generation.

The expression *intergenerational mobility* refers to the fact that parents' status does not determine the status of the child (when mobility is low, we say that the intergenerational transmission of wealth or any other measure of economic status is high). Economists and sociologists measure social mobility by ranking parents by their incomes or wealth, and then looking at what income or wealth their kids end up with when they become adults. High mobility implies that two kids from two different families, whose parents had different economic status, would both have the potential to succeed as adults. Knowing how rich the parents were would not tell us much about the likely income or wealth of the adult children. However, if mobility is low, then kids whose parents had a high income are likely to grow up to have high incomes themselves, and kids from low-income families are likely to have low incomes as adults.

This is what we see in Figure 19.12, which gives measures of the intergenerational mobility of men, based on their labour earnings in the US (top panel) and Denmark (the bottom panel). The tall bar on the left in the top panel means, among men whose fathers were in the bottom fifth of the earnings distribution, 40% were themselves in the poorest fifth, while 7% ended up in the top fifth of the earnings distribution. By contrast, 36% of the men born to the richest fifth were themselves in the richest fifth—the tall purple bar on the right.

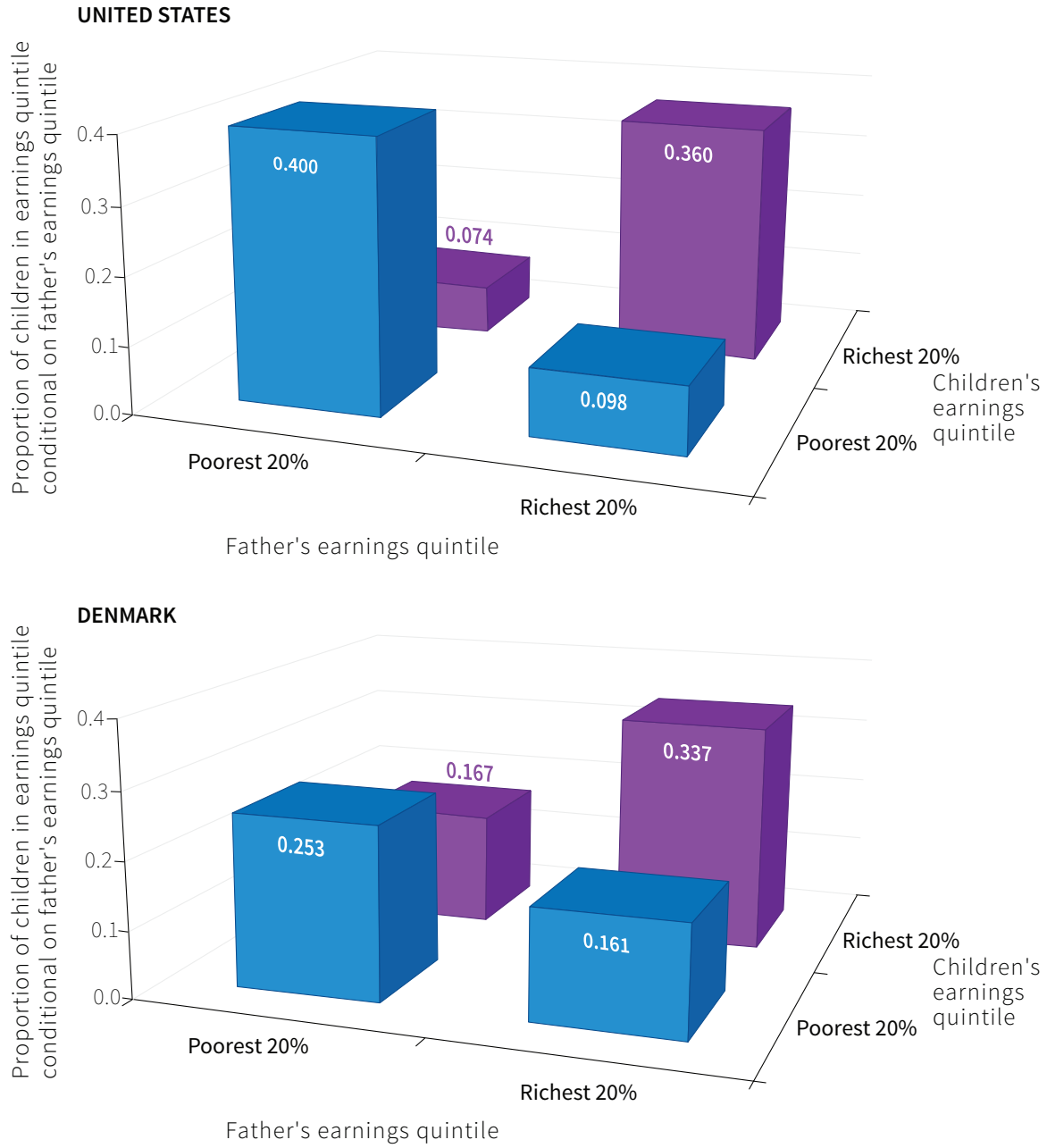


Figure 19.12 *Intergenerational transmission of earnings: US and Denmark.*

Source: *Table 14 in Jäntti, Markus, Bernt Bratsberg, Knut Røed, Oddbjørn Raaum, Robin Naylor, Eva Österbacka, Anders Björklund, and Tor Eriksson. 2006. American Exceptionalism in a New Light: A Comparison of Intergenerational Earnings Mobility in the Nordic Countries, the United Kingdom and the United States. Discussion Paper Series 1938. Institute for the Study of Labor.*

One of the reasons that children of the rich tend to be richer than the children of the poor is the financial support that rich parents give to their children, both during the parents' lifetimes and at death in the form of bequests. The data in Figure 19.12, however, is based on labour earnings, not inherited wealth. The earnings of parents and their children appear to be similar in the US, partly because children of well-off parents receive more, higher-quality, schooling. They also benefit from the networks and connections of their parents, which improve access to the labour market. High-earning parents may also pass on to their children the personalities and habits that

contributed to their high earnings. There may be other genetic explanations, but, other than race, it is difficult to document any genetically heritable traits that make a substantial contribution to intergenerational transmission of economic status. There is a significant genetic element in performance on cognitive tests such as IQ, for example, but this explains very little of the fact that parents and their offspring have similar incomes.

The data from Denmark in the lower panel suggests a more level playing field: the advantage of being born in a rich family is relatively small, and the disadvantage of being born poor is smaller than in the US.

A summary measure that tells us the overall rate of economic mobility in a society is called the intergenerational elasticity for some indicator of economic success such as wealth or income. To see what this measures, consider two pairs of fathers and sons: the father in the first pair is richer than the father in the second. The intergenerational elasticity measures how much richer the son of the richer father will be than the son of the poorer father. An elasticity of 0.5, for example, means that if one father is 10% richer, then his son on average will be 5% richer than the other son. The higher the intergenerational elasticity, the lower the mobility in the society, and the greater the degree of intergenerational transmission of economic status. We use this elasticity to measure what we term intergenerational inequality as distinct from what we usually refer to as inequality, which is measured between people, or groups of people, at a particular time.

INTERGENERATIONAL MOBILITY, ELASTICITY, INEQUALITY

Be careful not to confuse these three terms:

- *Intergenerational mobility*: Changes in the relative economic or social status between parents and children. When this is high, the parents' status does not determine the status of the child.
- *Intergenerational elasticity*: For parents and grown offspring, the percentage difference in the second generation's status that is associated with a 1% difference in the adult generation's status.
- *Intergenerational inequality*: The extent to which differences in parental generations are passed on to the next generation, as measured by the intergenerational elasticity or the intergenerational correlation.

Figure 19.13 presents evidence on the intergenerational elasticity for earnings and earnings inequality at a particular time, which is called cross-sectional inequality. Cross-sectional inequality is measured using the Gini coefficient for income. Note that we do not include the effects of taxes and government transfers in Figure 19.13 when we measure both income inequality and intergenerational transmission of

earnings, because we are interested in how these two dimensions of inequality move together independent of government policy. You can see the differences between measures of inequality in earnings, income and disposable income for a few countries in Figure 19.6.

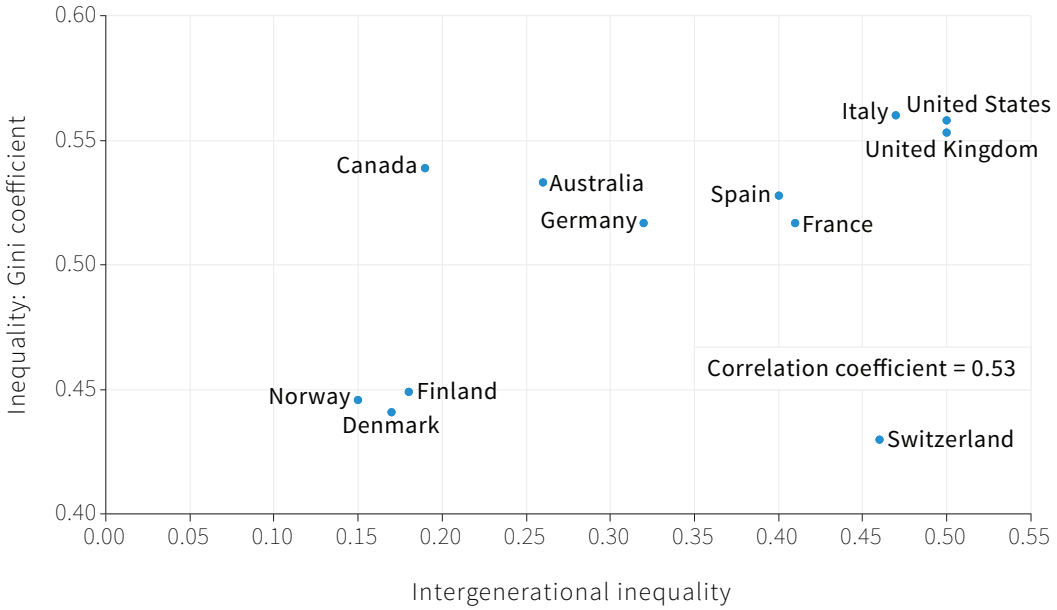


Figure 19.13 *Intergenerational and cross-sectional inequality.*

Source: *Corak, Miles. 2013. 'Inequality from Generation to Generation: The United States in Comparison.' In The Economics of Inequality, Poverty, and Discrimination in the 21st Century, edited by Robert S Rycroft. Santa Barbara, CA: Greenwood Pub Group; Chen, Wen-Hao, Michael Förster, and Ana Llena-Nozal. 2013. Globalisation, Technological Progress and Changes in Regulations and Institutions: Which Impact on the Rise of Earnings Inequality in OECD Countries? Working Paper Series 597. LIS.*

The figure shows that, for the countries considered, inequality in earnings at a given point is greater where high-earning fathers have high-earning sons, and the sons of low-earning fathers also receive low wages or salaries. But countries differ in which type of equality is most pronounced: compare, for example, Canada and Switzerland.

Does inequality cause intergenerational transmission of economic status, or the other way around, or both, or neither? We know that societies with a strong culture of fairness and equal treatment such as Denmark adopt policies to reduce inequality among people at a given moment, and at the same time attempt to improve intergenerational mobility by providing equal opportunities for high-quality education, by reducing the importance of inherited wealth, and through other policies that would reduce intergenerational transmission of economic status. We assume this is part of the explanation of the contrast between Denmark and the US in Figure 19.12.

Another likely source of the correlation in Figure 19.13 is that, in any period (a generation for example), some individuals experience good luck and others bad, for example from serious illness of themselves or a family member, unplanned

pregnancy, business failure or because technological change or shifts in demand make their skills less valuable. Where intergenerational inequality is high, for example in the US, Italy and the UK, high or low earnings resulting from good luck or bad luck are passed on to the next generation, and added to whatever shocks the next generation experiences. As a result, intergenerational inequality contributes to inequality at a specific moment.

19.7 WHAT (IF ANYTHING) IS WRONG WITH INEQUALITY?

Michael Norton, a professor of business administration, and Daniel Ariely, a psychologist and behavioural economist, asked a large sample of Americans how they thought the wealth of the US should be distributed: what fraction of it, for example, should go to the top 20%? They also asked them to estimate what they thought the distribution of wealth actually was. Figure 19.14 gives the results, with the top three bars showing the distribution that different groups of respondents considered would be ideal, and the fourth bar the wealth distribution that they thought actually existed in the US. This bar (labelled "Estimated") shows that they thought that the richest 20% owned about 60% of the wealth. The bottom bar shows the actual distribution. In reality the richest fifth owns 85% of the wealth. Even more surprising than the public's underestimate of wealth inequality is what they think a fair distribution should be. The top bar shows that Americans thought that, ideally, the richest 20% should own a little more than 30% of total wealth.

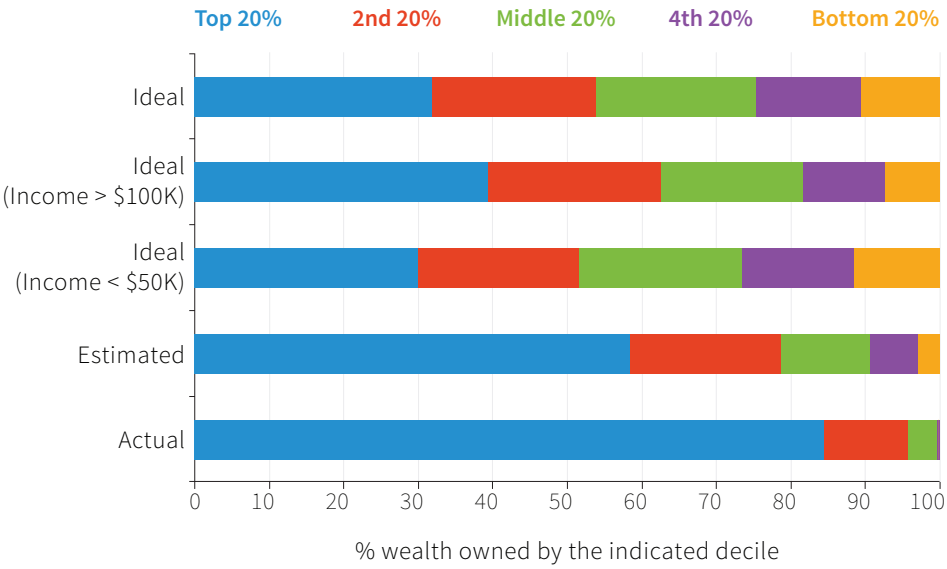


Figure 19.14 *Americans' ideal, estimated and actual distribution of wealth.*

Source: *Adapted from Figures 2 and 3 in Norton, Michael I, and Daniel Ariely. 2011. 'Building a Better America--One Wealth Quintile at a Time.' Perspectives on Psychological Science 6 (1): 9–12.*

Different groups largely agree on the ideal distribution of wealth. Americans with an annual income greater than $100,000 thought that the share going to the top 20% should be slightly larger than those who earned less than $50,000 thought it should be; Democratic party voters wished for a more equal distribution than Republican party voters; and women preferred more equality then men did. The differences between these groups, however, were small.

DISCUSS 19.6: ESTIMATED, IDEAL AND ACTUAL DISTRIBUTIONS OF WEALTH

Use this Gini coefficient calculator to determine the Gini coefficients for wealth ownership given by the estimated, ideal, and actual distributions in Figure 19.14. (Note: you will have to estimate the data visually from the chart.)

Although there seems to be a consensus on the ideal outcome in the US, policies that would redistribute income are controversial and debated passionately—as they are in most countries. Differences in self-interest contribute: richer Americans, for example, tend to oppose redistribution that favours the poor, while poorer Americans support it. But, as the experiments in Unit 4 would lead us to expect, self-interest is just part of the explanation. People differ also because they hold different beliefs about why the poor are poor and how the rich became rich. In laboratory settings, people often express strong feelings of fairness, and give up considerable sums of money to ensure outcomes that are consistent with ideas of economic justice. For example, responders in the ultimatum game reject what they consider an unfair offer, preferring to receive nothing and to impose the same fate on the proposer than to agree to being treated unfairly. Both rich and poor may think that high levels of inequality are unfair and that the government should reduce economic disparities, even if it means voting for policies that would reduce the disposable income of the voter.

Christina Fong, an economist, found that a person who thinks that hard work and risk-taking are essential to economic success is much less likely to support redistribution to the poor than one who thinks that the key to success is inheritance, race, your connections, or who your parents are. The results of her study are in Figure 19.15. Notice that white people who think that race (more precisely, being white) is important to getting ahead strongly support redistribution to the poor—evidently because they think that the process by which economic success is determined is unfair.

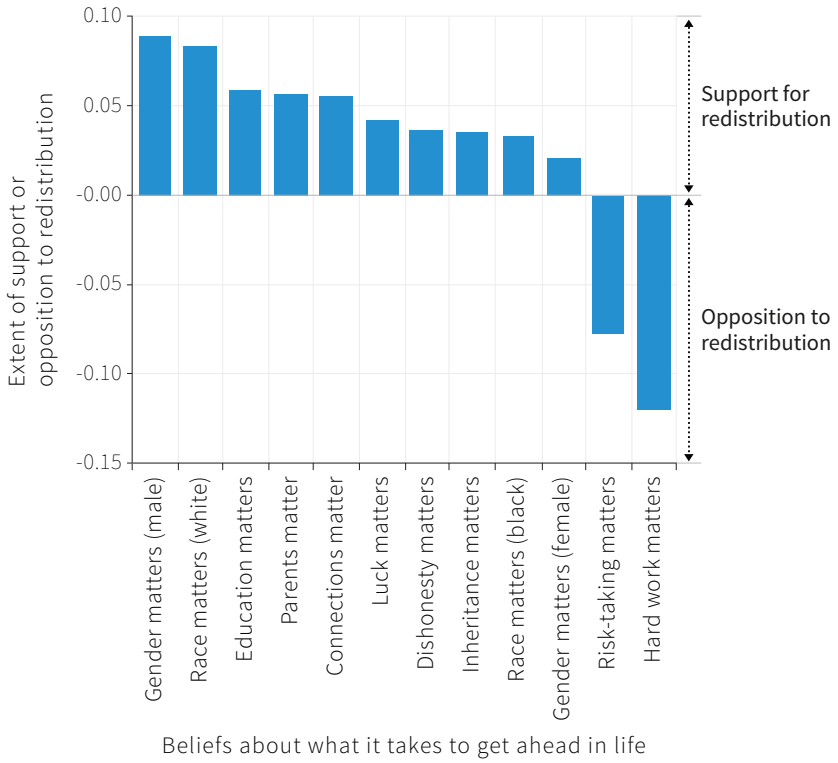


Figure 19.15 *How Americans' beliefs about what it takes to get ahead predict their support or opposition to government programs to redistribute income to the poor.*

Source: *Figure 5.3 in Bowles, Samuel. 2012. The New Economics of Inequality and Redistribution. Cambridge: Cambridge University Press.*

This suggests that the question "how much inequality is too much?" cannot be answered unless we know why a family or person is rich or poor. If income depends substantially on what we call an "accident of birth"—your race, your sex or your country—many people think this group inequality is unfair. We saw large group inequalities based on where you are born in Figure 19.1: have another look.

DISCUSS 19.7: A LEVEL PLAYING FIELD

When people think about "too much inequality", some think about the Gini coefficient measuring inequality at a point in time, while others are more interested in intergenerational inequality.

1. Use an example of two fictional families in each country to explain the combination of cross-sectional and intergenerational income inequality in Canada and Switzerland in Figure 19.13.
2. Rank the countries in Figure 19.13 according to your own values: which do you think is the fairest, the second fairest, and so on?

Now think about the indifference curves that you could draw in this figure that would indicate the combinations of inequality and intergenerational inequality that would be equally fair in your judgement.

3. If you cared only about a low Gini coefficient, what would they look like?
4. If you cared only about the intergenerational elasticity, what would your indifference curves look like?
5. If you cared about both, what would they look like, and why?
6. Use your rankings of the countries to draw a set of your own indifference curves (the curves implied by the ranking you have given in part 2, above).
7. Did drawing your own indifference curves lead you to revise your rankings?

19.8 HOW MUCH INEQUALITY IS TOO MUCH (OR TOO LITTLE?)

We have considered a range of policies that affect the degree of inequality and we will consider some more policies below. When we evaluate these policies, we ask two questions:

- What is the level of economic inequality that we would like our society to have?
- What are policies that will implement that level of inequality for the lowest cost?

This is exactly what we did in Unit 18. Given the trade-offs we face, we first determined the level of environmental quality that we would like to achieve, and then considered which policies would best implement the abatement that this implies.

The veil of ignorance as a lens for viewing inequality

Suppose your income depends substantially on what we call an *accident of birth*, whether that be your race, your sex, a physical disability or your country. Many people would regard any substantial differences in income between people arising from accidents of birth as unfair. People would be less likely to think of inequality as unfair if there was a *level playing field*, so that economic success is entirely a result of hard work.

This is what Fong's research in Figure 19.15 showed: those who think that "hard work matters" for getting ahead were opposed to the government redistributing income to the poor, while those who think that parents, connections, or inheritance matter held the opposite view.

But even if the playing field were level, we still would face a question: how rich should the winners be compared to the losers?

To think about this question, transport yourself to a hypothetical world in which you (perhaps along with other fellow citizens) are asked to design your model society. There will be two classes of equal size, one called "richer" and the other "poorer". You will get to live in the society you design after you have answered the question "how rich should the richer class be and how poor should the poorer class be?"

But there is a hitch: which class you get to be in will be determined by the flip of a coin.

This thought experiment is what the American philosopher John Rawls, who we encountered in Unit 5, termed choosing a social contract from behind a *veil of ignorance*: we would not know which positions we would occupy in the society we were considering.

Behind the curious device of the veil is an important concept: Rawls' fundamental idea is that justice should be impartial. It should not favour one group over another, and the veil of ignorance invites you to think this way (because you do not yet know which group you are going to be in). Rawls asked us to think about justice if:

"[N]o one knows his place in society, his class position or social status; nor does he know his fortune in the distribution of natural assets and abilities, his intelligence and strength, and the like."
– John Rawls, *A Theory of Justice* (1971)

This does not tell us the answer to how much inequality there should be, but it does suggest a way to look at it.

Feasible inequality

Is this question just about philosophical statements of our values, or can economics help? Economics is useful, because it gives us tools for studying what combinations of the income of the rich and the poor are feasible, and how we might reason about which ones are preferable to others.

Let's try one way to answer "how rich should the richer class be and how poor should the poorer class be?": let's say there should be no difference between the incomes of the rich and the poor. Suppose that, in this case, both classes would receive $100,000 annually (per adult). This is shown by point *E* in Figure 19.16 where the 45-degree line gives all of the points of equal income among the two classes. The figure shows the annual income per adult of the poor and the rich along the axes.

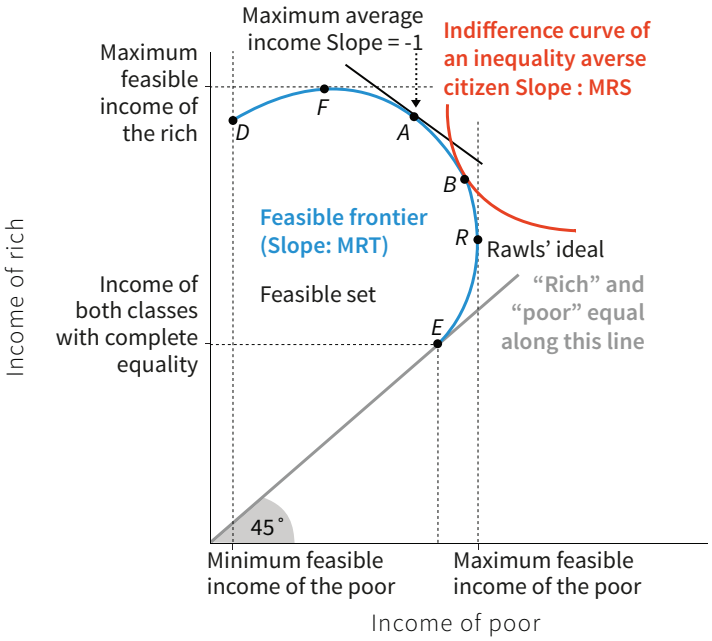


Figure 19.16 *Choice among feasible income distributions.*

Would this be your choice? In this version of an ideal society, you would not run the risk of ending up poorer than others after the coin flip. But as an economist you might think that complete equality in the society would mean that there were insufficient incentives for people to work and study and take risks, so that at least a little bit of inequality could actually be better for everyone.

In the figure points between *E* and *R* show possible combinations, in which the rich are richer than the poor, but where the poor also are richer than they would be under complete equality. Comparing the two points you can see that *E* is Pareto-inefficient because both rich and poor are better off at *R* than at *E*. The income distribution at *R* is also the one at which the poor are as rich as they can possibly be in the economy in question. This is the point that Rawls favoured (and why we called it point *R*).

The upward sloping part of the line between *E* and *R* is similar to a part of the feasible environment-consumption frontier in Figure 18.27 in which at very low levels of abatement, increased environmental quality is possible along with increased consumption.

Would you choose *R*? Notice that above *R*, the frontier is very steep. This means that it's possible to make the rich richer at very little reduction in the income of the poor. The blue curved line made passing through *R* and *E* (and the other points above *R*) is the frontier of the feasible set of income distributions for the economy in question. As with all feasible frontiers, the slope is a marginal rate of transformation, in this case transformation of income losses of the poor into income gains for the rich.

$$\text{slope of the feasible frontier} = MRT = \frac{\text{income gains for the rich}}{\text{income losses for the poor}}$$

If the point *R* had been proposed, would you want to consider other points higher up on the feasible frontier? Remember, after the coin flip, you will get either the income of the rich or that of the poor with equal probability (one half), so you know that:

$$\text{expected income} = 0.5 \times (\text{income of the rich}) + 0.5 \times (\text{income of the poor})$$

So as long as income gains for the rich come at little expense of income losses for the poor, you would definitely do better to move above point *R*. If you were interested in maximising your expected income then you would choose point *A*, where the income gains of the rich are exactly offset by income losses by the poor, so the marginal rate of transformation is equal to one.

But, after point *A*, the inequalities would become so severe that the average income would fall, and the rich would be getting a larger slice of a smaller pie. This might occur if the poor were insufficiently well fed to work hard, or sufficiently angry about their condition to motivate the rich to employ people to guard their gated communities, rather than producing goods and services. By looking ahead to Figure 19.19c, you will see data showing that more unequal societies (such as the US, UK and Italy) devote more resources to workers employed in private and public security activities than do other more equal countries with similar GDP per capita.

You can also see that if you could rig the coin flip so that you knew you would end up rich (and you had no concern about fairness) you would select point *F*.

Like the feasible set when Angela and Bruno were bargaining, there is a minimum level of income that the poor can get. This minimum could be set by their biological needs, or perhaps by the fact that were income to fall below this level they would revolt. Notice that if the poor were to be even poorer than at point *F*, the rich would themselves suffer. So like point *E* (maximal equality), point *D* (minimal income of the poor) is not Pareto-efficient.

In the figure we have considered the following income distributions:

- *E*: complete equality
- *R*: the distribution with the highest income for the poor
- *A*: the highest average income of rich and poor
- *F*: the maximum income of the rich
- *D*: the distribution in which the poor are at their minimum feasible living standard

A preference for fairness

Which would you choose? Points between D and F are easy to eliminate from the running as they are all inferior for both classes to point F. And the same goes for points between E and R. The same idea—eliminate from consideration all Pareto-inefficient distributions—means that no points interior to the feasible set would be considered: like points between D and F and between E and R, they are all Pareto-inefficient.

That leaves points between F and R. How will you choose among them? To answer this, you need to consult your indifference curves. In this case an indifference curve gives combinations of the incomes of the two classes that are equally valued by you. Curves further away from the origin are preferred (more income for both is always better).

The slope of these indifference curves is the marginal rate of substitution between income for the rich and income for the poor.

$$\begin{aligned} \textit{slope of the feasible frontier} &= MRS \\ &= \frac{\textit{marginal value of poor income}}{\textit{marginal value of rich income}} \end{aligned}$$

You would then maximise your utility by finding the point on the feasible frontier at which the marginal rate of transformation is equal to the marginal rate of substitution. If you wished to maximise your own expected income, then you would place an equal value on the income of the rich and the poor because you are equally likely to be one or the other.

But you might care about the condition of the poorer class even were you to have the good luck to be assigned to the richer class in the coin flip (remember you have to make your choice before you know your assignment). That is, you might be *inequality averse*, caring about your own payoffs but also disliking inequality per se. In this case you would have an indifference curve like the red one shown in the figure. You would choose point B, somewhere between Rawls' ideal (the highest feasible income of the poor) and point A, the highest average income.

19.9 ADDRESSING UNFAIR INEQUALITY

Where inequalities are seen as unfair, government policies can limit economic inequality in four ways:

- *Using taxes and transfers*: They reduce disparities in disposable income.

- *Reducing individual differences in endowments*: Maybe reduce the difference in wealth, or the determinants of success in the labour market.
- *Change the relative value of endowments*: Adopt policies to increase the value of endowments held by poorer people.
- *Insuring citizens against some economic losses*: This would reduce the impact of bad luck, such as job loss or ill health, on economic status (see Unit 6 to remind yourself about this).

Using taxes and transfers

In Figure 19.5b we saw that inequality in wealth and before-tax labour earnings is much greater than the inequality of disposable income—the buying power of a family after paying taxes and receiving any government transfers such as unemployment insurance. The equalising effect of taxes and transfers is particularly high in Sweden (where wealth is unequally held, compared to most countries) but lower in Japan (where both earnings and wealth are not very unequally distributed, at least compared to the US). We also saw in Unit 1 that some other countries (Belgium and Germany) create low levels of income inequality using taxes and transfers.

Giving money directly to poor people is one possible transfer:

- *Conditional transfers*: In Latin America, governments give money to poor households that comply with certain actions such as sending their kids to school (measured by attendance records) or having them vaccinated. Conditional cash transfer programs such as Oportunidades in Mexico and Bolsa Familia in Brazil have gained support from middle class voters, and transferred substantial income to the poor. Economic researchers have found that Oportunidades has yielded long-term improvements in children's schooling; children have entered the workforce later and are more likely to be in non-agricultural work.
- *Unconditional transfers*: In South Africa, cash transfers given as pensions have had significant effects on child nutrition and schooling achievement because the pension goes primarily to the grandmother, who is often the person responsible for the care and schooling of her grandchildren. We know that unconditional cash transfers like this increase the autonomy poor people have in deciding what to do with their resources, and are relatively cheap to administer.

Greater equality of endowments

Other countries (South Korea and Taiwan for example) are more like Japan, experiencing limited levels of inequality without a major role of taxes and transfers because endowments are more equally distributed. After the second world war, all three redistributed the property of large landowners among landless or land-poor farmers and have also invested heavily in providing high-quality education to citizens.

These countries have focused on redistributing endowments, rather than taxing well-off citizens to raise the incomes of the less well-off through transfers from the government or adopting policies to raise the value of endowments held by poor people.

James Heckman, a Nobel prizewinning economist, shows how economists can learn from experiments and other data how to level the playing field for children growing up poor in our Economist in Action video.

Raising the value of the endowments of the poor

Governments can also intervene in market processes directly by creating, protecting or breaking up monopolies or by employing large numbers of workers directly. Using the legal system, governments can also alter which property rights are protected, for example banning slavery, establishing trading rights in emissions (Unit 18), or setting the duration of intellectual property rights and patents. All of these measures can change which groups have bargaining power and their reservation options, which will change the distribution of income. Finally, governments can also change the set of contracts that are allowed or enforced, and this alters the distribution of the income as we saw in Unit 5, when we saw the effect of legislation to enforce maximum hours of work. Another example is a *statutory minimum wage*, which prohibits contracts with wages below a certain level.

Figure 19.17 shows how a minimum wage can affect the Lorenz curve and the Gini coefficient. We begin with the segmented labour markets economy in which half the workers, those in the secondary labour market, make just one-third of what the primary segment workers earn (the solid Lorenz curve). Now we consider a minimum wage that doubles the wages in the secondary market. The immediate impact of this change is that the 40 workers making up the secondary labour market now receive 20%, rather than 10%, of the total income. The result is that the Gini coefficient falls from 0.62 to 0.49. To assess the longer-run effects of the minimum wage we would have to consider the effects on the profits of their employer, the greater spending of the secondary labour market workers and other indirect effects on the level of employment.

MINIMUM WAGE

The statutory minimum wage is a tool used in advanced economies:

- It eliminates labour contracts where wages are too low
- It reduces inequality between the middle and the bottom of the income distribution

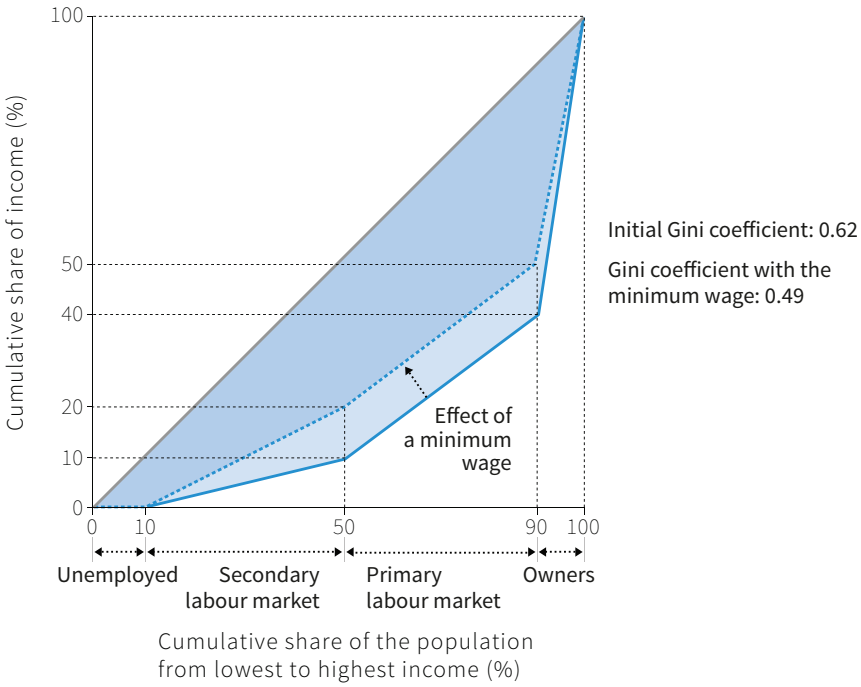


Figure 19.17 *The effect of a minimum wage in a segmented labour market economy.*

The costs of the minimum wage could be unemployment, although recent evidence suggests that this is small for minimum wage rises in the US and UK. A study of different changes in minimum wages in areas in the US, when that area was next to a similar location in which the minimum wage did not rise, showed that raising the wage more than offset the decrease in employment. Therefore, on average, raising the minimum wage increased the income of poor workers.

Legislation may also rule out particular kinds of contracts, such as those that prohibit employees from leaving their firm to work for a competitor. The justification offered for these *non-compete contracts* is that workers leaving a firm may take with them industrial or trade secrets that would benefit the competition. But in the US non-compete contracts are now being imposed even on fast food workers. This suggests employers have a different motive: to reduce the reservation option of employees by making it more difficult to find work should they be fired. This lowers the wage. Or, to put it in policy terms: prohibiting non-compete contracts would raise the wage.

A third policy is illustrated by the Indian state of West Bengal where a tenancy reform, Operation Barga (discussed in Unit 5) gave local bargadars (sharecroppers) the right to keep three-quarters of their crop rather than handing over half of the crop to the landowner, as had been the previous custom. They also received protection from eviction by landowners as well as rights to their product.

19.10 EQUALITY AND ECONOMIC PERFORMANCE

The success of Operation Barga in raising productivity in farming, of Oportunidades in Mexico, and of pensions in South Africa in raising school achievement and child health may help explain a fact that some people find surprising: more equal countries do as well, or better, in terms of standard economic performance than unequal countries.

We saw in Unit 17 (Figure 17.15) that the low levels of inequality, the power of trade unions, and the growth of pro-poor tax and transfer policies during the golden age of capitalism were associated with the most rapid growth of income per capita in modern history. Investment, too, occurred at levels not seen before, raising the capital stock at an unprecedented rate of growth.

Earlier in this unit (Figure 19.8) we showed the centuries-long U-turn of top incomes in many countries including the US and the UK. By this measure, late 20th-century inequality had risen to levels not experienced since before the Great Depression. But this U-turn pattern is far from universal, as Figure 19.18 shows.

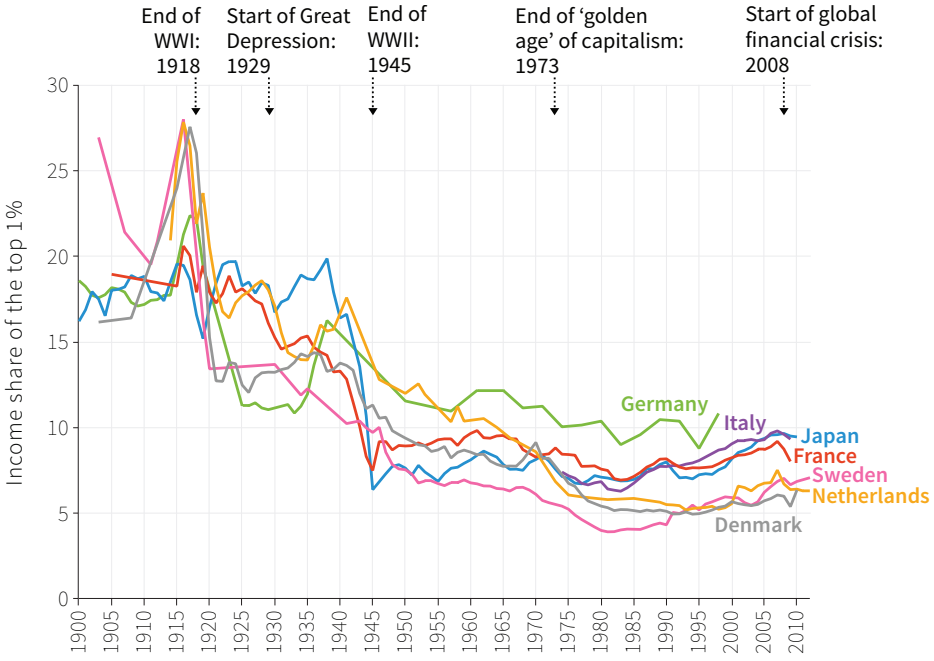


Figure 19.18 *Declining share of the top 1% in some European economies and Japan.*

Source: *Alvaredo, Facundo, Anthony B Atkinson, Thomas Piketty, Emmanuel Saez, and Gabriel Zucman. 2016. 'The World Wealth and Income Database (WID).'*

DISCUSS 19.8: THE U-TURN COUNTRIES

Look again at the difference between the U-turn countries in Figure 19.8, which showed a trend towards greater equality in the first three quarters of the 20th century followed by an increase in inequality since about 1980, and the countries in Figure 19.18, in which inequality did not increase significantly, or at all.

Make a list of possible explanations as to why countries in the two groups took such different courses since 1980, making sure to check (you can use the internet) that any technological or institutional changes you refer to are historically accurate.

Most of the countries in Figure 19.18 are high performers when we look at countries that achieved both rapid growth in income per capita and modest levels of inequality of disposable income, as you can see in Figure 19.19a. We measure inequality in income after taxes and transfers (disposable income) here because this is the best measure of the inequality in living standards that people experience. Differences between measures of inequality in earnings, income and disposable income can be seen for a few countries in Figure 19.5b. Figure 1.14 in Unit 1 shows that, in some countries, government policies result in disposable income being considerably less unequal than income. The most evident conclusion from Figure 19.19a is that most countries grow at similar rates and there is no relation to the level of inequality.

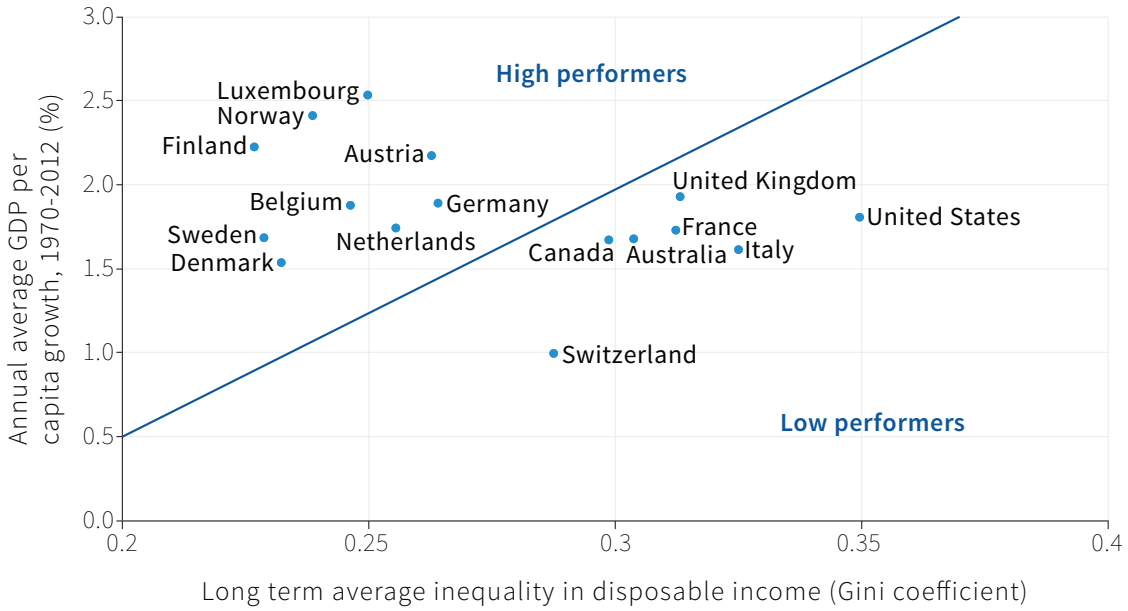


Figure 19.19a *The cost of inequality: Inequality and growth in living standards among rich countries.*

Source: *Wang, Chen, and Koen Caminada. 2011. 'Leiden Budget Incidence Fiscal Redistribution Dataset.' Version 1. Leiden Department of Economics Research.*

DISCUSS 19.9: HIGH AND LOW PERFORMERS

We have drawn a line in Figure 19.19a to distinguish high from low performers, but what counts as "high" performance depends on your preferences.

1. Rank the countries in Figure 19.19a from the most to the least preferable from your standpoint (for example, would you prefer less inequality and a slower rate of growth as in Switzerland, or more inequality and a higher rate of growth as in the US?).
2. Use your rankings to sketch your indifference curves in the space given by Figure 19.19a (hint: is the slope of the indifference curve positive or negative?).

There have also been high and low performers among the catch-up countries. Figure 19.19b shows that South Korea and Taiwan were able to achieve high growth with relatively low inequality over the past 30 years, whereas the performance of Latin American economies along these dimensions was typically much worse.

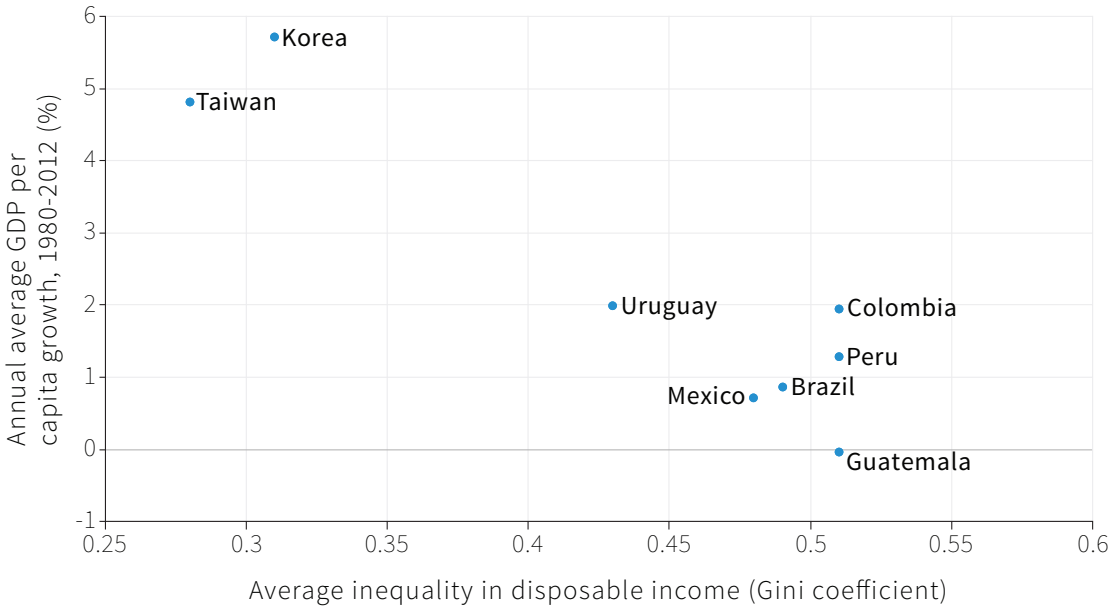


Figure 19.19b *The cost of inequality: Inequality and growth in living standards among catch-up countries.*

Source: *Wang, Chen, and Koen Caminada. 2011. 'Leiden Budget Incidence Fiscal Redistribution Dataset.' Version 1. Leiden Department of Economics Research; OECD; International Monetary Fund. 2014. 'World Economic Outlook Database: October 2014.'*

Figures 19.19a and 19.19b are initially surprising because economists have often claimed that high taxes and transfers depress incentives for people to work hard and take the kinds of risk necessary for innovation to occur. Explanations of why egalitarian countries such as Japan, South Korea, and Taiwan in Asia, and Nordic and other northern European countries have done so well economically include:

- *Cooperation and trust*: this helps to create high-quality, knowledge-intensive goods, but it may be difficult to sustain when inequalities within a firm or a nation are large.
- *Policies that enhance the endowments of the poor*: High-quality health services, education, and land ownership by the person who works on the land contribute to the more productive use of an economy's resources. This is also true of policies that raise the value of the endowments of the poor, as illustrated by Operation Barga.
- *Guard labour*: Nations characterised by high levels of inequality divert a substantial portion of their resources away from productive uses and into the construction of secure environments for the rich, such as gated communities.

Figure 19.19c illustrates this last point: the US, Italy and the UK are countries with highly unequal disposable incomes that hire three times as many guards (public and private security personnel, excluding armed forces) than do the more equal nations of Finland, Denmark and Sweden. An unequal society may expend a lot of resources on protecting property rights and enforcing the rule of law. More unequal societies may have more of their workforce devoted to repressing, persuading, or otherwise managing potential conflicts.

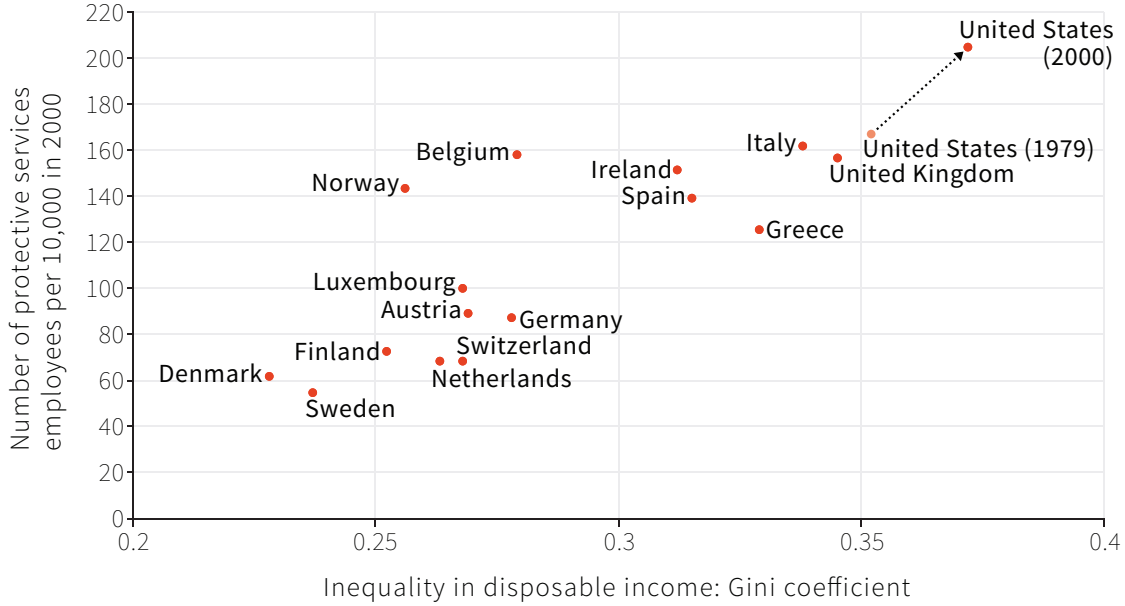


Figure 19.19c *The cost of inequality: Economic disparity and the fraction of workers employed as guards.*

Source: *Jayadev, Arjun, and Samuel Bowles. 2006. 'Guard Labor.' Journal of Development Economics 79 (2): 328–48.*

19.11 CONCLUSION

What have we learned about the people that we represented by the hypothetical Mark, Yichen, Stephanie and Renfu, and how their economic lives turned out differently?

Their starting conditions were shaped by their families, which set them up with differing initial upbringing, skills and physical wealth (endowments) and opportunities. The scarcity or abundance of those endowments on markets and hence the value of their endowments then shaped their opportunities. Globalisation (trade between China and the US) and technology (the Titan robot and the Motorola plant) reduced the value of Mark's skills as a machinist. China—US trade had the opposite effect on Yichen's fortunes: her factory job at Motorola paid much more than she ever would have made as a farmer, and it would not have existed had the plant not exported its products to the US.

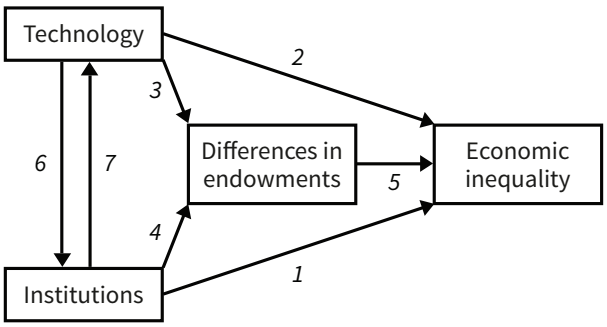
Differences in the endowments that the four developed in schooling (university against high school), as well as institutions (unions, taxes, and hukou passports), all played a role in driving their lifetime incomes. Economists interested in the study of inequality spend a lot their time trying to untangle the interactions and relative importance of these forces.

We can summarise much of this Unit, and earlier ideas about inequality, using the model in Figure 19.20, adding examples of each of the arrows, now numbered in the figure. Think about arrow number 1: this is a direct channel from institutions to economic inequality exemplified by:

- *Greater competition in markets*: Opening of the economy to more imports will reduce the markup on costs, and redistribute income from owners of firms to consumers, making the endowments of the owners less valuable than before.
- *Minimum wages and immigration*: A minimum wage raises the value of the endowments of some workers, immigration lowers it.

Direct effects of technology on economic inequality (arrow 2) include labour-saving innovations that reduce demand for, and hence the value of, the endowments of workers with skills that can be replaced by machines.

Figure 19.20 has examples for the other numbered arrows.



Arrows	Example	Units
1	The degree of competition in markets, bargaining power of employers and employees, taxes, government transfers, property rights, provision of public goods, change in immigration policy affects value of one's nationality or one's skill, minimum wage, tariffs	10, 16, 7
2	Displacement of workers by labour saving innovations, new technologies that require more skills or increase the productivity of some endowments	2, 15
3	A new technology may make a previously useless skill a source of income (e.g. in code-writing); or remove a handicap (such as blindness) from limiting sources of income, or country of residence a source of income (e.g. analysing radiology results remotely)	2, 20
4	Inheritance tax, land reform, educational opportunities, intellectual property rights, rule of law, role of force in acquiring endowments	5, 10
5	Changes in supply and demand affect rates of return on endowments	7, 8
6	Digital revolution affects intellectual property rights, extent of increasing returns affects degree of competition in markets	20
7	Intellectual property rights, degree of competition among firms, education, FDI policy, rule of law	20

Figure 19.20 *Institutions, technologies and differences in endowments as causes of economic inequality. The numbered arrows in the top panel refer to the numbered rows in the bottom panel.*

But there is something missing from Figure 19.20: change.

You know from Unit 1 that a capitalist economy is constantly changing and changing the world. From Figures 19.7, 19.8, 19.9 and 19.18 in this unit you know that the level of wealth and income inequality changes over time. This is partly a result of the changes in technology and in institutions due to the dynamism of capitalism.

In Figure 19.21 we extend the model to make it less like a snapshot and more like a film. Think of the earlier figure as a frame in a film. When you run the film you see that the story does not end with the three arrows representing effects on economic inequality. Economic inequality is both an effect of other influences and a cause of institutions, technologies and endowments in the future.

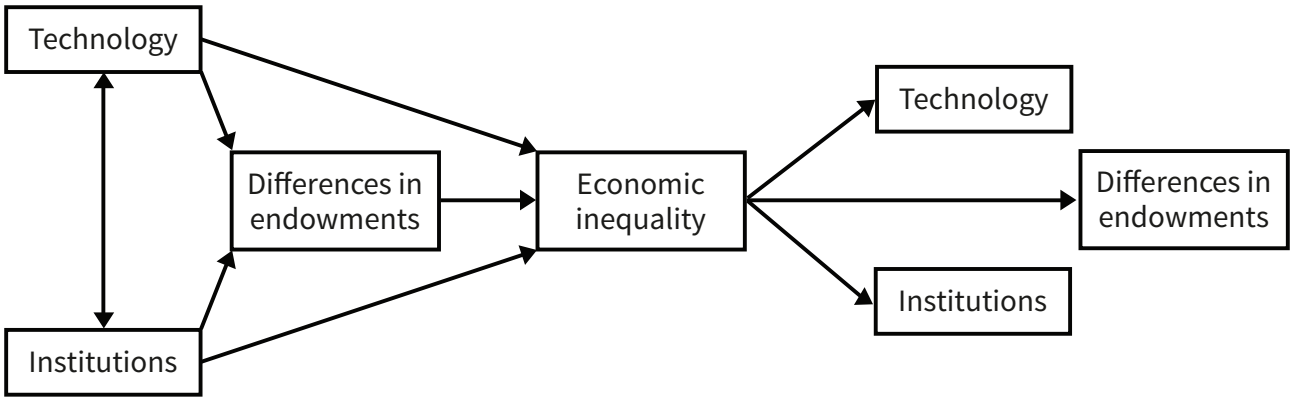


Figure 19.21 Determinants of inequality in this and future generations.

Figure 19.12 (the earnings of fathers and their children in the US and Denmark) showed that although economic status is not automatically transmitted to the next generation, the earnings of fathers helps predict the earnings of children. It's a much better predictor in some countries than others, as we saw in Figure 19.13.

A high level of inequality may, even in a democratic society, allow the rich to affect laws that are adopted. In this way they control the institutions that shape the lives of the next generation. An example is laws that weaken trade unions. On the other hand, a lower level of inequality may produce institutional change in the school system that makes the playing field more level for the next generation of children.

CONCEPTS INTRODUCED IN UNIT 19

Before you move on, review these definitions:

- *Gini coefficient*
- *Lorenz curve*
- *Endowment*
- *Technology*
- *Institution*
- *Classes*
- *Labour market segmentation*
- *Group inequality*
- *Intergenerational elasticity*
- *Inequality aversion*
- *Minimum wage*
- *Labour market segmentation*

The arrow from inequality to technology captures the way in which high inequality creates market incentives for innovation activities to be focused more on serving the cosmetic needs or extending the longevity of the rich than on the needs of the poor. In the next unit we continue this theme of the dynamism of the capitalist economy, focusing on innovation and technical change in the production and distribution of knowledge. Among other questions, we will look into the possible effect on inequality of new technologies capable of replacing human cognition with artificial intelligence, and substituting robotic systems for human labour.

Key points in Unit 19

Lorenz curve, Gini coefficient, intergenerational elasticity

The Lorenz curve and the Gini coefficient provide summary measures of the degree of inequality; the intergenerational elasticity is a measure of the transmission of economic success from parents to children.

Technology and institutions influence inequality

Both influence the level of inequality directly, and also through their effects on inequalities in endowments.

Credit and labour markets, and wealth, influence inequality

Markets in credit and labour, alongside the distribution of wealth, influence inequality in income by affecting the relationships among classes including lenders and borrowers (and the credit-excluded), and employers and employees (and those without work).

Global inequality

For the world population, most inequalities in income today are the result of accidents of birth, the most important of which are one's nation, parents, gender, or ethnic or cultural group.

Government policies to reduce inequality

Policy can reduce inequalities by affecting technology, institutions and the distribution of endowments.

How much inequality?

Judgements about how much inequality a society should have involve trade-offs that can be analysed using a feasible frontier of combinations of inequality and average income, with indifference curves for values such as inequality aversion and self-interest.

The optimum level of inequality

When we do this type of analysis, many economies appear to be well inside the feasible frontier.

Costs and benefits of economic equality

Policies to reduce inequality may have adverse effects on incentives. But countries and historical periods characterised by greater economic equality have also experienced higher growth of productivity and living standards, and low unemployment.

19.12 EINSTEIN

Inequality as differences among people

The Gini coefficient, g, is often calculated as the area between the Lorenz curve and the complete equality line divided by the area under the equality line. This, however, is an approximation (see the note on small population bias, below). It is more precisely defined mathematically as *half of the relative mean difference in incomes among all pairs of individuals in the population*. We can calculate g easily using this definition when we know the income of every person in a small population.

So, to calculate g when we know the incomes of every member of a population:

1. We find the difference in income between every possible pair in the population
2. Then we take the mean of these numbers
3. We divide this number by the mean income of the population. This statistic is the *relative mean difference*
4. g = *relative mean difference divided by two*

Thus, if there are just two individuals in the population and one has all of the income, we can assume their incomes are 0 and 1:

1. The difference between the incomes of the pair = 1
2. This will be the mean difference because there is just one pair
3. Mean income = 0.5, so the relative mean difference = 1/0.5 = 2
4. $g = 2/2 = 1$ (as we would expect)

Now think about two people dividing a pie. The Gini coefficient is a measure of how unequal their slices are. For example:

1. If the smaller slice of the pie is 20% of the pie, the other person's slice is 80%, so the difference is 60% (0.60)
2. This is the mean difference (there are only two incomes, as before)
3. Divide this by the mean slice size (which is 50% or 0.50), to get 1.20
4. Half of this is 0.60, which is the Gini that measures the inequality between the two pie eaters when the unfortunate one gets only 20% of the pie

In the illustration below, this division of the pie is the third picture. In each case we show the Gini coefficient when two people share a pie:

g = 0 **g = 0.2** **g = 0.6** **g = 1**

Now suppose there are three people, one of whom has all of the income, which we assume is an income of 1 unit.

1. The differences for the three possible pairs would be 1, 1 and 0
2. The mean difference = 2/3
3. The relative mean difference = (2/3)/(1/3) = 2
4. The Gini coefficient $g = 2/2 = 1$

Again, $g = 1$, as we would expect.

As an exercise, show that in the two-person case, where the size of the smaller slice is σ, we can write:

$$\sigma = \frac{1 - g}{2}$$

(Hint: rearrange the equation so that g is on the left-hand side.)

To make sure you understand the four steps in this method, take the three-person case and show that $g = 0$ if the incomes are equal, and calculate g if one person has zero, another has 1/3 and the third person has 2/3.

The Lorenz curve, the Gini coefficient and a small population bias

When considering a population with small numbers (like the dividing-the-pie example), the Lorenz curve method does not give a good approximation of g, especially if there is high inequality. To see why, try drawing a Lorenz curve for a population of just two individuals in which the poorer person (50% of the population) has no income and the richer person has all the income. In this case we have just seen that $g = 1$. But calculating the Gini coefficient from the areas in the Lorenz diagram you get $g = 0.5$.

Start with the Lorenz curve expression for the Gini coefficient:

$$g = \frac{A}{A+B}$$

Using this expression, we see that in the two-person case, where one person has all the income:

$$g = \frac{A}{A+B} = 0.5$$

But, if one person has all of the income, clearly $g = 1$. When N is the size of the population, and N is small, the corrected Gini coefficient calculated using the Lorenz curve method is:

$$g = \left(\frac{N}{N-1}\right)\left(\frac{A}{A+B}\right)$$

In this case, $N = 2$:

$$g = \left(\frac{2}{2-1}\right) \times 0.5 = 1$$

So the Gini coefficient, corrected for small population size, equals one (as it should).

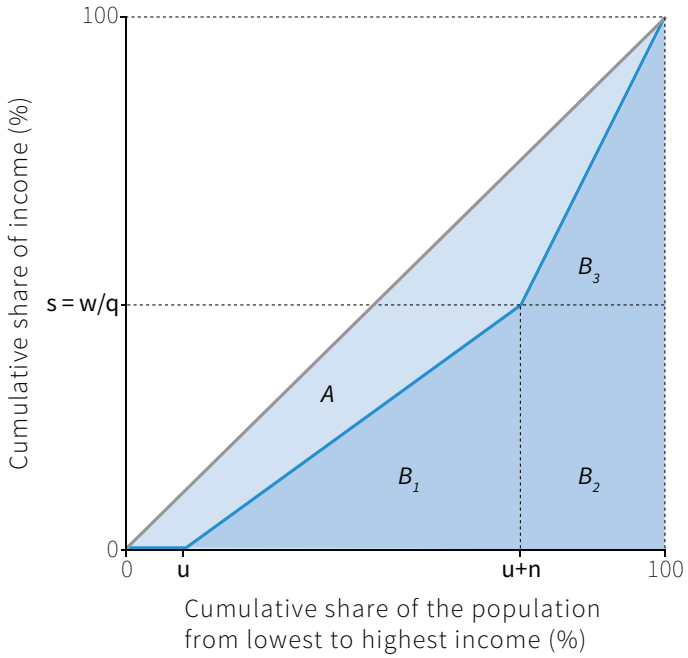
To summarise:

- The definition of the Gini as half the relative mean differences is always correct irrespective of population size
- It is approximated by the ratio of the areas A/(A+B) in the Lorenz curve diagram (such as Figure 19.5a) if the population is large.

The Lorenz curve and the Gini coefficient in a class-divided economy with a large population

A population normalised to 1 is composed of the following fractions: u is unemployed, n employed, $(1\text{-}n\text{-}u)$ is made up of employers or landlords, or other claimants on income who are not producers. We consider a large population so that we can approximate the Gini coefficient by the expression:

$$g = \frac{A}{A+B}$$



Output is average product of the producers (q) times the number of producers (n). It is composed of the following fractions: w/q received by the workers (that, is, the producers) and $1\text{-}w/q$ received by their employers (or other non-producers). Total output (nq) is normalised in the figure to be unity. The share of output received by the producers is $s = w/q$.

You can see from the figure that the slope of the line dividing region A from region B_1 is:

$$slope = \frac{s}{n} = \frac{1}{n} \times \frac{w}{q}$$

In the Lorenz curve figure shown, letting $B = B_1 + B_2 + B_3$, the Gini coefficient is:

$$g = \frac{A}{A+B_1+B_2+B_3}$$

We now use the figure to derive an equation for the value of the Gini coefficient in terms of the variables:

- u, the fraction of the population that is unemployed
- n, the fraction of the population that is employed (or more generally produce output)
- w, the real wage
- q, output per employed worker (producer)

Our strategy is to express the areas of the two triangles and one rectangle that make up region B in terms of the above variables, and then subtract that expression from 0.5 (which is the total A+B area under the line of complete equality). We start with the areas below the Lorenz curve:

$$B_1 = \tfrac{1}{2}\, nw/q$$
$$B_2 = (1 - u - n)w/q$$
$$B_3 = \tfrac{1}{2}\,(1 - u - n)(1 - w/q)$$

To find B:

$$\begin{aligned} B &= B_1 + \tfrac{1}{2} B_2 + \tfrac{1}{2} B_2 + B_3 \\ B &= \tfrac{1}{2} nw/q \\ &\quad + \tfrac{1}{2}(1 - u - n)w/q \\ &\quad + \tfrac{1}{2}(1 - u - n)w/q \\ &\quad + \tfrac{1}{2}(1 - u - n)(1 - w/q) \\ B &= \tfrac{1}{2}(1 - u)w/q + \tfrac{1}{2}(1 - u - n) \\ &= \tfrac{1}{2}\{(1 - u - n) + (1 - u)w/q\} \end{aligned}$$

Which means we can derive an expression for g:

$$\begin{aligned} g &= \frac{A}{A + B} \\ &= \frac{1/2 - B}{1/2} \\ &= 1 - \frac{B}{1/2} = 1 - 2B \\ &= u + n - (1 - u)w/q \end{aligned}$$

What can we learn from the expression $g = u + n - (1 - u)w/q$?

- If the non-producer class (for example, employers or landlords) gets smaller (relatively), that is, as u or n increase (holding constant the share of income of the non-producer claimants), inequality goes up, as one would expect. This could depict the early evolution of capitalism from an economy of smallish family-owned firms and manufacturers employing a few workers to a modern economy, with concentrated wealth and hence few non-producing claimants.
- An increase in the wage share (or share of total output retained by the producers if we have a sharecropping economy) *ceteris paribus*, will reduce the Gini.
- If the identical producers keep all that they produce ($w/q = 1$), then there are no non-producing income claimants, and if there is no unemployment, then $g = 0$.
- In the heading we specified a large population. If the population is small, the Gini coefficient calculated this way does not equal 1 when a single person receives all of the income. To show this, suppose that $w = 0$, so the only income goes to the employers, $g = u + n$. Now imagine that there are 10 people in the population, just one of whom is the employer. Then $g = 0.9$. This is the small population bias, as we saw in the section above. If you calculate the Gini coefficient by taking differences among pairs of people in the population, your result for g will not be subject to this small population bias.

This Gini coefficient equation can help us understand inequality in past and present societies:

An extreme class society

- A large population with a single employer: In this case, $u+ n \approx 1$ and, using this approximation, g = 1 - (1 - u)w/q. If there are no unemployed: g = 1 - w/q. This is the profit share, for example the landlord's crop share. So, in this case, *the Gini coefficient is entirely determined by the class relationship of producer to non-producer*.
- *A large population of producers with a single member of the elite*: Let's call him the king. The king taxes the producers. Then *the Gini coefficient is the tax rate*, that is, the share of output that the producers hand over to the king.

Demographic segmentation and robot production

- *There are no non-producer income claimants*: If producers receive their entire output then $w/q = 1$ and $u + n = 1$, so $g = u$: the *Gini coefficient is the rate of unemployment* (fraction of population not engaged as producers). Here the Gini coefficient is entirely determined by demographic structure, that is the fraction of the population who are producers.
- *Robotic systems take over from humans*: As a result, 60% of the labour force is without work, and without pre-tax and transfer income. Then even if the employed workers were to receive a pretax wage equal to the amount they produce (so there would be no "owners") the Gini coefficient would be 0.6, making *this hypothetical economy more unequal than most of the countries in the world today* (see Figure 1.16).

19.13 READ MORE

Bibliography

1. Acemoglu, Daron, and James A. Robinson. 2012. *Why Nations Fail: The Origins of Power, Prosperity, and Poverty*. New York, NY: Crown Publishing Group.
2. Alvaredo, Facundo, Anthony B Atkinson, Thomas Piketty, Emmanuel Saez, and Gabriel Zucman. 2016. 'The World Wealth and Income Database (WID).'
3. Atkinson, Anthony B. 2015. *Inequality: What Can Be Done?* Cambridge, MA: Harvard University Press.
4. Atkinson, Anthony B, and Thomas Piketty, eds. 2007. *Top Incomes over the Twentieth Century: A Contrast between Continental European and English-Speaking Countries*. Oxford: Oxford University Press.

5. Berg, Andrew G, and Jonathan D Ostry. 2011. Inequality and Unsustainable Growth: Two Sides of the Same Coin? *Staff Discussion Note SDN/11/08*. International Monetary Fund.
6. Bertrand, Marianne, and Sendhil Mullainathan. 2004. 'Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination.' *American Economic Review* 94 (4): 991–1013.
7. Bourguignon, Francois. 2015. *The Globalization of Inequality*. Princeton, NJ: Princeton University Press.
8. Bowles, Samuel. 2012. *The New Economics of Inequality and Redistribution*. Cambridge: Cambridge University Press.
9. Bowles, Samuel, and Arjun Jayadev. 2014. 'One Nation under Guard.' *New York Times*, February 15.
10. Bowles, Samuel, and Herbert Gintis. 2002. 'The Inheritance of Inequality.' *Journal of Economic Perspectives* 16 (3): 3–30.
11. Carnevale, Anthony P, Stephen J Rose, and Ban Cheah. 2011. *The College Payoff*. Georgetown University Center on Education and the Workforce.
12. Chen, Wen-Hao, Michael Förster, and Ana Llena-Nozal. 2013. *Globalisation, Technological Progress and Changes in Regulations and Institutions: Which Impact on the Rise of Earnings Inequality in OECD Countries?* Working Paper Series 597. LIS.
13. Clark, Gregory. 2015. *The Son Also Rises: Surnames and the History of Social Mobility*. Princeton, NJ: Princeton University Press.
14. Corak, Miles. 2013. 'Inequality from Generation to Generation: The United States in Comparison.' In *The Economics of Inequality, Poverty, and Discrimination in the 21st Century*, edited by Robert S Rycroft. Santa Barbara, CA: Greenwood Pub Group.
15. Daly, Mary C, and Leila Bengali. 2014. 'Is It Still Worth Going to College?' Federal Reserve Bank of San Francisco. May 5.
16. Deaton, Angus. 2013. *The Great Escape: Health, Wealth, and the Origins of Inequality*. Princeton, NJ: Princeton University Press.
17. Diamond, Jared. 1999. *Guns, Germs, and Steel: The Fates of Human Societies*. New York, NY: Norton, W. W. & Company.
18. Dube, Arindrajit, T. William Lester, and Michael Reich. 2010. "Minimum Wage Effects across State Borders: Estimates Using Contiguous Counties." *Review of Economics and Statistics* 92 (4): 945–64.
19. Flannery, Kent, and Joyce Marcus. 2014. *The Creation of Inequality: How Our Prehistoric Ancestors Set the Stage for Monarchy, Slavery, and Empire*. Cambridge, MA: Harvard University Press.
20. Fochesato, Mattia, and Samuel Bowles. 2013. *Technology, Institutions and Wealth Inequality in the Very Long Run*. Santa Fe Institute.
21. Fochesato, Mattia, and Samuel Bowles. 2013. 'Wealth Inequality from Prehistory to the Present: Data, Sources and Methods.' Dynamics of Wealth Inequality Project, Behavioral Sciences Program, Santa Fe Institute.
22. Heckman, James J. 2013. *Giving Kids a Fair Chance*. Cambridge, MA: MIT Press.

23. International Monetary Fund. 2014. 'World Economic Outlook Database: October 2014.'
24. Jayadev, Arjun, and Samuel Bowles. 2006. 'Guard Labor.' *Journal of Development Economics* 79 (2): 328–48.
25. Jäntti, Markus, Bernt Bratsberg, Knut Røed, Oddbjørn Raaum, Robin Naylor, Eva Österbacka, Anders Björklund, and Tor Eriksson. 2006. *American Exceptionalism in a New Light: A Comparison of Intergenerational Earnings Mobility in the Nordic Countries, the United Kingdom and the United States*. Discussion Paper Series 1938. Institute for the Study of Labor.
26. Milanovic, Branko. 2007. *Worlds Apart: Measuring International and Global Inequality*. Princeton, NJ: Princeton University Press.
27. Milanovic, Branko. 2012. 'Global Income Inequality by the Numbers: In History and Now -an Overview-.' *Policy Research Working Papers*. The World Bank.
28. Milanovic, Branko. 2012. *The Haves and the Have-Nots: A Brief and Idiosyncratic History of Global Inequality*. New York, NY: Basic Books.
29. Norton, Michael I, and Daniel Ariely. 2011. 'Building a Better America--One Wealth Quintile at a Time.' *Perspectives on Psychological Science* 6(1): 9–12.
30. Piketty, Thomas. 2014. *Capital in the Twenty-First Century*. Cambridge, MA: Harvard University Press.
31. Rawls, John. (1971) 2009. *A Theory of Justice*. Cambridge, MA: Belknap Press of Harvard University Press.
32. Sutcliffe, Robert B. 2001. *100 Ways of Seeing an Unequal World*. London: Zed Books.
33. The World Bank. 2016. 'IIASA/VID Educational Attainment Model. Dataset Produced by the International Institute for Applied Systems Analysis (IIASA) in Laxenburg, Austria and the Vienna Institute of Demography, Austrian Academy of Sciences.' *Educational Attainment Statistics*.
34. Waldenström, Daniel, and Jesper Roine. 2014. 'Long Run Trends in the Distribution of Income and Wealth.' In *Handbook of Income Distribution: Volume 2a*, edited by Anthony Atkinson and Francois Bourguignon. Amsterdam: North-Holland. Data.
35. Wang, Chen, and Koen Caminada. 2011. 'Leiden Budget Incidence Fiscal Redistribution Dataset.' Version 1. Leiden Department of Economics Research.

Beta June 2016 version

INNOVATION, INFORMATION, AND THE NETWORKED ECONOMY

INNOVATIONS THAT ENHANCE OUR WELLBEING ARE A HALLMARK OF CAPITALISM. MAKING THE MOST OF HUMAN CREATIVITY AND INVENTIVENESS IS A CHALLENGE TO PUBLIC POLICY

- The process of innovation is influenced by the state of knowledge, individual creativity, public policy, economic institutions, and social norms that jointly make up an innovation system
- When the institutions of the capitalist economy are working properly, successful innovators gain innovation rents, which are eventually competed away by imitators
- As a result, the amount of knowledge and the number of innovations have dramatically increased since the advent of capitalism
- In economics, knowledge is unusual in two ways: it is a public good, and its production and use are characterised by extraordinary economies of scale
- Economies of scale and means of delaying imitation (such as patents) mean that knowledge-producing firms are at least temporary monopolists, and can profit from winner-take-all competition by setting prices above the marginal costs of production
- Innovating firms cannot capture all of the benefits their innovations will generate, because the knowledge they produce is a public good
- Public policy can be designed so that innovations are used more quickly, by more people
- Public policy seeks to encourage others to copy successful innovations while at the same time providing adequate rewards for inventors. Therefore intellectual property rights can be either “too strong”, deterring copying, or “too weak”, providing innovation rents that are too small to motivate potential inventors
- If public policy gets this trade-off right, continued innovations may create a future that is more environmentally sustainable, less unequal, and one in which we may be able to work less

Institute for New Economic Thinking
UCL SciencesPo Azim Premji University
FRIENDS PROVIDENT Foundation

South Africa has long had one of the highest rates of HIV infection in the world: about 5 million South Africans, one in 10 of the population, are HIV positive. But in 1998, Bristol-Myers Squib, Merck and 37 other multinational pharmaceutical companies brought a lawsuit against the government of South Africa, seeking to prevent it from importing generic (non-brand name) and other inexpensive antiretroviral drugs and other HIV/AIDS treatments from around the world.

Street protests erupted in South Africa, and both the European Union and the World Health Organization announced their support for the South African government's position. In September 1999, the US government—previously the drug companies' strongest ally—said that it would not impose sanctions on poor countries ravaged by HIV/AIDS even if US patent laws were broken, so long as the countries abided by international treaties governing intellectual property. The pharmaceutical giants pushed back, engaging virtually every intellectual property rights lawyer in the country to promote their case. They closed factories in South Africa and cancelled planned investments.

But three years later, with millions of dollars spent on litigation and with the even greater cost to their reputations, the companies backed down (even paying the South African government's legal fees). Jean-Pierre Garnier, the chief executive officer of GlaxoSmithKline, telephoned Kofi Annan, secretary general of the United Nations, to ask him to help make a deal with Thabo Mbeki, the president of South Africa. "We're not insensitive to public opinion. That is a factor in our decision-making," Garnier explained.

It was too late: the damage had already been done. ''This has been a public relations disaster for the companies,'' commented Hemant Shah, an industry analyst. ''The probability of any drug company suing a developing country on a life-saving medicine is now extremely low based on what they learned in South Africa.''

Probably true; but put yourself in the shoes of the owners of the pharmaceutical companies. They also worry about survival. They cannot sell an HIV/AIDS treatment at what it cost them to manufacture it and stay in business. Moreover, few of the industry's research projects lead to a marketable product (maybe one in 20, although this is a hard number to estimate). The sales of a successful product must therefore cover the costs of 19 failed projects because, of course, it is impossible to predict which research projects will succeed.

This is why the drug companies went to court in South Africa to protect their patents. In the pharmaceutical industry, the patent system gives the innovating company a time-limited monopoly on the product that allows the company to charge a high price (above manufacturing cost) during the years of patent protection. The prospect of high profits provides an incentive for companies to invest in risky research and development.

In the Economist in action video, F.M. Scherer, an economist who specialises in the effects of technological change, explains how patents support innovation in pharmaceuticals.

Patent protection, by creating a government-imposed monopoly, often conflicts with the equally important objective of making goods and services available at their marginal cost. (Recall from Unit 7 that a monopoly will set a price above the marginal cost.) The high price—sufficient to cover the cost of research and development, including on failed projects—means that many of those who could benefit from access to the drug will not get it.

Conflicts between competing objectives—in this case the production of new knowledge on the one hand and its rapid diffusion on the other—are unavoidable in the economy, and are particularly difficult to resolve when they concern innovation, as we will see.

But sometimes new technologies allow for win-win outcomes.

Recall the problem of the fishermen and fish buyers of Kerala that we described at the beginning of Unit 9. When returning to port to sell their daily catch of sardines to fish dealers, fishermen often found that there was a glut on the market. They dumped their worthless catch back into the sea. Sometimes a lucky few returned to the right port at the right time when demand exceeded supply, and they secured high prices. The result was higher prices for the consumer, on average, and lower incomes for the fishermen.

This all changed when the fishermen got mobile phones. While still many kilometres out to sea, the returning fishermen would phone the many coastal fish markets, and pick the one where the prices that day were highest. By gaining access to real-time market information on relative prices for fish, the fishermen could adjust their pattern of production and distribution to secure the highest returns. A study of 15 beach markets along 225km of the northern Kerala coast found that, after the fishermen began using mobile phones, differences in daily prices among the beach markets were a quarter of their previous levels. No boats dumped their catch. Reduced waste and the elimination of the fish dealers' bargaining power raised the profits of fishermen by 8% while consumer prices fell by 4%.

The cell phone implemented the *Law of one price* in Kerala fish markets, to the benefit of fishermen and consumers. It was not entirely win-win, however. The mobile phones greatly increased the competition among the dealers who were the intermediaries between fishermen and buyers, because a fisherman could bargain for higher prices before choosing which market to enter.

The mobile phone revolution had much weaker effects in other parts of the world, such as Uttar Pradesh and Rajasthan in India, where lack of roads and storage facilities prevented farmers from profiting from information on price differences. A small farmer in Allahabad remarked that price information that he could get on his

phone was not worth much to him because there were "no roads to go there." In this case the innovation was of little use, because of a lack of public investment in the necessary infrastructure.

Similarly, when mobile phones came to Niger, in West Africa, it was mostly the traders, not the farmers, who benefited—in part because they also lacked the means to transport their cowpeas and other crops to alternative markets.

In this unit we will show how economic concepts can make sense of the South African government's policies to make HIV/AIDS treatments more widely available, the conflict that the policies caused, and the contrasting impacts of the mobile phone on fishermen in Kerala and farmers in other Indian states.

To understand innovation, you will have to forget about the image of an eccentric inventor, working alone, creating a better mousetrap, and getting rich as a reward for the public benefit of his inspiration. Innovation is not a one-off event set off by a spark of genius, but instead:

- *Innovation is a system:* It connects networks of users, private firms, individuals, and government bodies.
- *Innovation is also a process:* It is a fundamental source of change in our life that itself is constantly undergoing change.

20.1 INNOVATION: INVENTION AND DIFFUSION

We begin with a few new terms. We use the word *innovation* to refer to both the development of new methods of production and new products (*invention*) and the spread of the invention throughout the economy (*diffusion*). An innovating firm can produce a good or service at a cost lower than its competitors, or a new good at a cost that will attract buyers. The first is called a *process innovation* and the second is called a *product innovation*.

You already studied many of the concepts that are useful for the study of innovation. They are listed in Figure 20.1, and you will encounter them again throughout this unit. Before going on, make sure you understand and recall these concepts.

INVENTION AND INNOVATION

The descriptive term *invention* is sometimes reserved for major breakthroughs, but we use it to refer to:

- *Radical innovation:* The invention of incandescent lighting (producing light by running electricity through a filament) was a major advance over light made by burning oil or kerosene. Radical innovations have been found to rely on a broad range of knowledge from different sectors, recombining this to create new and very different products. For instance, the inventors of the MP3 format relied on research into how humans perceive sound. They made sound files that were smaller because they understood which information could be dropped without altering what listeners would hear.
- *Incremental innovation:* This improves an existing product or process cumulatively. After Edison and Swan worked out the basic design of incandescent light bulbs in 1860, the subsequent improvements in the filament that generates the light were incremental innovations in lighting. You have already learned about the incremental improvement of the spinning jenny—one of the major inventions of the Industrial Revolution—which began with just eight spindles and eventually operated hundreds.

CONCEPTS	PREVIOUSLY IN UNITS
Innovation rents	1 and 2
External effects & public goods	10
Strategic interactions	4, 5, 6
Property rights including IPR	1, 2, 5, 10
Economies of scale	7
Complements and Substitutes	3
Mutual gains and conflicts over their distribution	5
Creative destruction	2, 15
Institutions and social norms	4, 5, 15

Figure 20.1 *Concepts relevant to innovation that you have studied.*

Recall from Units 2 and 10 that, at the going price, a company introducing a successful invention makes profits in excess of the profits that other firms make, termed *innovation rents*. In Figure 20.2, the costs of undertaking process or product innovation are shown with the temporary innovation rents (profits above the opportunity cost of capital) from a successful invention.

The prospect of these innovation rents then induces others to try to copy the invention. If they are successful, the temporary innovation rents are competed away and the initial innovator earns profits that just cover the opportunity cost of capital, so economic profit is zero. Late-comers are also eventually pushed to adopt the innovation because the falling prices that result when the new methods become widely adopted typically mean that sticking with the old technology is a recipe for bankruptcy. A firm that does not innovate will make profits that fail to cover the opportunity cost of capital, so it makes negative economic profits and may be forced out of business. The carrot-and-stick combination of the promise of rents from successful innovation and the threat of bankruptcy if firms fail to keep up with innovators has proved a powerful force in expanding consumption possibilities and raising living standards.

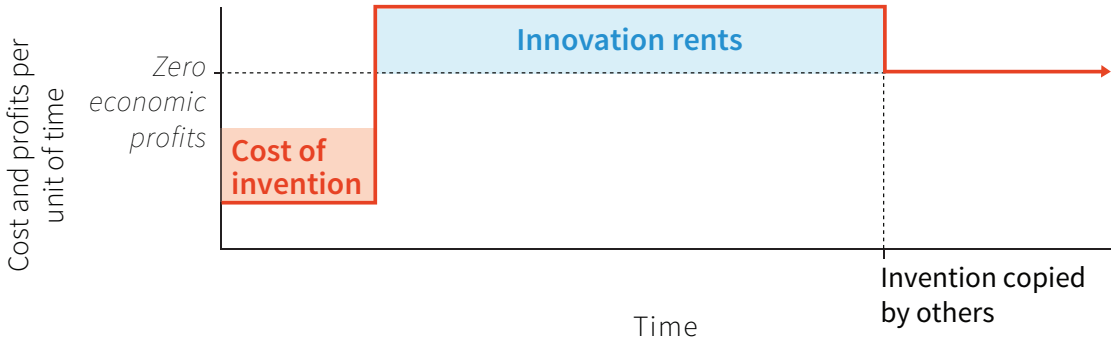


Figure 20.2 *The costs and rents associated with innovations.*

Although there have been inventions throughout human history, the process of capitalist innovation started in England around 1750 with some key new technologies introduced in textiles, energy and transportation. It did not end with the Industrial Revolution. Important new technologies with applications to many industries such as the steam engine, electricity, and transportation (canals, railroads, automobiles, airplanes) are called *general-purpose technologies*.

William Nordhaus, an economist whose analysis of the discount rate applied to environmental problems you read about in Unit 18, has estimated the speed of computation using an index which has a value of 1 for the speed of a computation done by hand (like dividing one number by another). For example, in 1920 a Japanese abacus master could perform computations 4.5 times faster than a mathematically competent person could do the same calculation by hand. This difference had probably been constant for many centuries, because the abacus is an ancient

computational device. But sometime around 1940, computational speed takes off. The IBM 1130 introduced in 1965 was 4,520 times faster than hand computation (and as you can see, it was below the line of best fit through the data points from 1920).

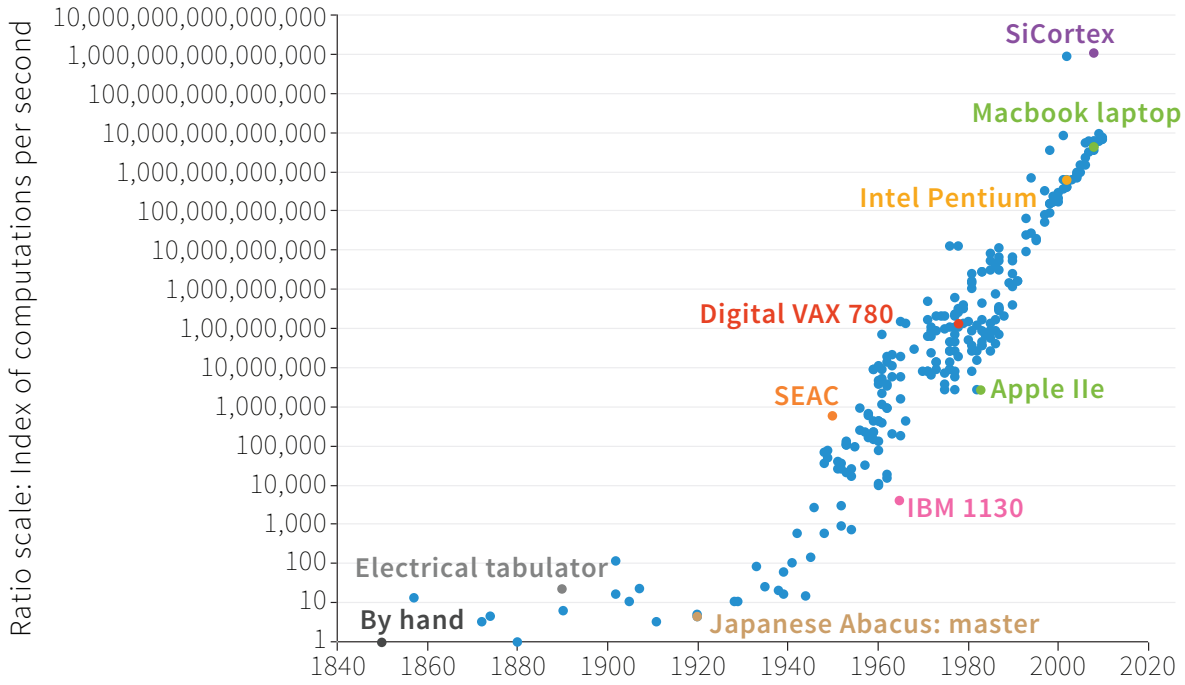


Figure 20.3 *Innovation in computing power: Index of computing speed. Particular examples are shown in colour and labelled.*

Source: *Nordhaus, William D. 2007. "Two Centuries of Productivity Growth in Computing." The Journal of Economic History 67 (01), Index updated to 2010.*

The most recent entry in Figure 20.3, the SiCortex supercomputer, performs 1 billion computations per second. It is more than a quadrillion (count the zeros) times faster than you, and it is well above a line of best fit through the data points from 1920, so there is no indication that the process is slowing down. But technologists—and economists—disagree over whether improvements in computation or any other technology will continue at the pace in Nordhaus's chart, or instead will return to the modest pace of improvement that prevailed over most of human history prior to the 20th century.

INNOVATION SYSTEM

A successful innovation system consists of two elements:

- Relationships among private firms, governments, educational institutions, individual scientists and other actors involved in the invention, modification, and diffusion of new technologies
- These social interactions are governed by a combination of laws, policies, knowledge, and social norms

The stepped line in Figure 20.2 illustrated a simple theory of innovation and diffusion of technical progress. It clarifies how innovation rents, costs of innovation and the copying of innovations are interrelated from the standpoint of a firm or individual that wants to develop a new product or process.

But to understand this process, we need to know how inventions actually happen, how the costs and rents are decided, and when the process of copying takes place. To do this we have to go beyond the point of view of the single firm in Figure 20.2 to understand innovation as the product of interactions among firms, the government, educational institutions and many other players in the *innovation system*.

20.2 INNOVATION SYSTEMS

Innovative activities are not spread evenly across the globe or even across a country. Think of the area now known as Silicon Valley in California, once a sleepy farming area centred on Santa Clara Valley. Silicon Valley got its nickname when high-growth firms in computing and semiconductor design moved in, later joined by innovators in biotech. In 2010, in a single US postal area (ZIP code 95054) in the centre of Silicon Valley, 20,000 patents were registered. Patent attorneys cluster in this part of Santa Clara. If this small area of 16.2km2 were a country, it would have ranked 17th in the world in patents in 2010.

The outpouring of patents from Silicon Valley is an example of the vast amount of what is termed *codified knowledge* created there. But much of the knowledge produced cannot be written down, or at least not exactly. This non-codifiable knowledge is termed *tacit knowledge*.

The difference between codified and tacit knowledge can be illustrated this way: a recipe for a cake can be written down, and so is codified knowledge, but being able to read the recipe and follow it exactly does not get you a reputation for being an outstanding cook; on the other hand, the tacit knowledge of an exceptional chef is not something that you can easily write in a book. A historical example of the importance of tacit knowledge: after the first world war and again after the second world war, German chemicals companies had their factories in Germany disassembled and their facilities in the US and UK expropriated. All that remained were key personnel. Using their know-how and experience, German companies nevertheless managed to resume dominant positions in some markets.

Silicon Valley is as famous for its tacit knowledge as it is for the patented codified knowledge. The extraordinary concentration of innovative businesses in Silicon Valley reflects the importance of external effects and public goods in the production

and application of new technologies. The two words "Silicon Valley" no longer just refer to a place; they now represent a particular way that innovation gets done. Silicon Valley has become associated with an innovation system.

As well as the legal institutions that protect codifiable knowledge and that govern how easily holders of tacit knowledge can move between firms, an innovation system includes financial institutions such as venture capital funds, banks or technology-oriented firms that will finance projects that seek to commercialise innovations.

Different countries provide quite different innovation systems that often co-evolve with industries in which they specialise. For example, radical innovation is more prevalent in the US, where labour can move easily between firms and venture capital is well-developed, and incremental innovation is more prevalent in Germany, where ties of workers to firms are stronger and finance for innovation comes from retained profits and banks rather than from venture capital.

The Silicon Valley innovation system

Why is innovation concentrated in Silicon Valley? Institutions and incentives reinforce each other to produce a radical innovation cluster. The Silicon Valley model is one of highly mobile entrepreneurs, investors and employees linked within a small geographical area by networks that include universities, and an essential role for both the government of the State of California and the US government. The system consists of:

Entrepreneurial innovation firms

Much innovation takes place in firms specialising in producing new methods or products (start-ups) rather than in existing firms that produce goods and services.

Universities

In a partnership that began early in the 1900s, two universities (University of California at Berkeley and Stanford University) work closely with firms to commercialise innovations. An industrial park was set up in 1951 at Stanford with major corporations like General Electric, IBM and Hewlett Packard.

Government

Military research in electronics and high energy physics was funded at the universities and in private firms in the area, starting in the run-up to the second world war. In the Cold War this continued with Lockheed Missiles and Space, the largest employer in the Valley. A change in the law in 1980, called the Bayh-Dole Act, enabled universities to gain ownership of their output and commercialise it even if the federal government had helped fund it. This brought private investors into the network.

Entrepreneurs

A social norm for high risk, high return behaviour, which has its origins in the speculators who flooded into California to mine for gold in the 19th century, encourages innovators to exchange a part of their business in return for investment capital. High tolerance for business failure sustains a culture of serial entrepreneurship: failed innovators can start again with a new idea. Entrepreneurs at an early stage will pitch their project to venture capital (VC) investors. When the VCs decide to invest and take a substantial ownership stake, usually for a period of 12 to 18 months, it creates strong incentives for the startup to grow rapidly and, if successful, means the VC investor can exit with a high rate of profit.

Industrial research centres

University, government and private R&D labs are co-located in the Valley. A recent arrival is the R&D centre of Walmart, the retail giant.

Venture capital funding

The funding model for startups is a rapid-paced cycle of pitching a new business idea to investors which is based on the commercialisation of an invention, followed by recruiting key employees (often with earnings linked to the value of the firm when it is sold), growth of market, and seeking more cash. Founders, investors and employees all understand that failure is likely. Funders still benefit, because the few successful ventures produce large returns that compensate for many losses.

Relationships among firms

High firm failure rates and other reasons for employee mobility across firms distribute the tacit knowledge acquired in one firm to other firms. Some analysts have concluded that this unintentional sharing of information among firms was key to Silicon Valley's success. During the 1960s Silicon Valley was a minor player in technology compared to the Route 128 concentration in Massachusetts, benefiting from proximity to Harvard and MIT. As a way of protecting information that a firm produced Massachusetts enforced *non-compete contracts* that prohibited anyone leaving one firm from taking up employment with a competing firm. The law in California took the opposite position: "every contract by which anyone is restrained from engaging in a lawful profession, trade, or business of any kind is... void." The circulation of engineers among firms in Silicon Valley promoted the rapid diffusion of new knowledge among firms.

The German innovation system

Innovation in the US is concentrated in industries whose patents heavily cite scientific articles. This is one indicator of radical innovation. By contrast, the very successful export industries of Germany rely on incremental innovation where patents are much less intensive in scientific citations and tacit knowledge tends to be more important. Networks are crucial to the German innovation system but

they work differently from those in Silicon Valley. Like Silicon Valley, innovation is concentrated geographically, with centres around Munich and Stuttgart in southwest Germany.

Innovation in long-lived goods-producing companies

Incremental innovation takes place in medium-size and large long-lived companies in Germany, and relies on long-term relationships between employers and workers, between firms and banks, and among firms linked through both production relationships and ownership and control ties. To succeed in introducing new technology, firms face many coordination problems, and both cooperative and competitive relationships with employees, other firms and banks help solve them.

Employees

Skilled workers are needed for the successful introduction of process and product innovations. The machine tool, transport equipment and other capital goods industries that are innovation leaders in Germany face these problems:

- *For young people to commit to multi-year apprenticeships at low wages:* They need to be assured of long-term high wage employment.
- *For workers to engage in labour displacing innovation:* They must be assured that they will not lose their jobs as a result.
- *For firms to invest in costly training:* They need to know that their skilled workers will not be poached by other firms.

The German incentive system, and especially its vocational training aspect, addresses these problems in a variety of ways.

- *Social norms against poaching:* The assurance that firms' highly trained workers will not be poached is assured not by law but by norms that are widely respected by the otherwise highly competitive firms.
- *Government-subsidised apprenticeship system:* This reduces training costs for firms and ensures high quality training, which is supervised by industry associations. Apprentices contribute by accepting low training wages.
- *Certification:* This assures trainees that their skills are valuable outside the firm, improving their reservation position should their job end, and helping to ensure high wages as long as the job continues.
- *Works councils:* Large firms are required to have elected bodies to represent workers in negotiations with managers. They help to devise ways to exploit all possible mutual gains and to distribute these gains in a way acceptable to all.

Interactions among firms

Incremental innovation (for example, in the automobile industry) requires industry-wide standards to make technology transfer easier. Long-term relationships and cross-ownership among firms make this possible. The system of ownership of large

German firms differs sharply from that of US or UK firms. Takeovers are easier in the US or the UK, and allow for rapid changes in the use of the assets of a firm. Because ownership of firms is much more concentrated in Germany, it is virtually impossible for a hostile takeover—that is, one opposed by the management—to occur. Therefore, long-run inter-firm collaboration over technology development is possible, and industry-wide standards are easier to set. This is essential because long-term employment contracts mean that the Silicon Valley-style technology transfer that occurs when workers move from one firm to another is much less common.

Finance

The financing of innovation in Germany comes from retained profits (profits not distributed to shareholders) and bank loans. Long-term finance provides reassurance for trainees who invest in acquiring company-specific skills, as well as for other companies investing in related technology developments.

Figure 20.4 compares the two systems. Both are successful, but in different ways. Silicon Valley-based firms dominate important digital technologies (ICT) associated with the latest general-purpose technology, while the German firms making up its distinctive innovation system have managed to sustain a much higher level of well-paying industrial jobs in the face of global competition than has the US or any other country outside of East Asia.

DISCUSS 20.1: COMPARING INNOVATION SYSTEMS

In the text we compared the German innovation system with the system of innovation in Silicon Valley.

1. Do you think there are characteristics of the two systems that would make them incompatible within the same geographical and legal sphere?
2. On the basis of the above, can you imagine a clone of Silicon Valley in Germany, or a clone of the German manufacturing system in California (for example, for the manufacture of electric cars)? Explain your answer.

	SILICON VALLEY	GERMAN INNOVATION SYSTEM
INNOVATING FIRMS	Entrepreneurial innovation specialists	Established industrial and other firms
GOVERNMENT	Military contracts, higher education	Subsidies for training workers
INNOVATION	Radical codified, especially in ICT	Incremental tacit, especially in capital goods and transport equipment
INNOVATORS	Engineers, scientists	Skilled workers and engineers
PROPERTY RIGHTS	Patents of more importance	Non patent forms of protection of more importance
FINANCE	Venture capital	Bank loans, retained earnings
SOCIAL NORMS	Competitive; risk taking	Cooperative; risk pooling

Figure 20.4 *Two innovation systems: Silicon Valley and Germany*

The economics of innovation systems

Successful innovation can contribute to rising living standards by expanding the set of products available to consumers, and by reducing the prices of existing products. But many societies struggle to innovate. Compare the amount of innovation in capitalist economies to the amount in centrally planned economies of the Soviet Union and its allies during the 20th century: in a list of 111 major nonmilitary product and process innovations between 1917 and 1998, only one—synthetic rubber—came from Soviet bloc countries. Scholars have suggested that an important factor contributing to the collapse of the Soviet planned economies was the erosion of legitimacy of Communist Party rule because it failed to deliver innovation in consumer goods.

The successful capitalist innovation systems in Silicon Valley and Germany have two things in common:

- *The innovation system is not based on individual creativity:* A single firm or an inventor relies on the relationships among all of the actors—owners, employees, governments, and sources of finance. Regions without these support networks are less successful at innovation.
- *There is an invisible hand and a guiding hand:* Successful innovation combines profit-seeking competition among individuals and firms with government intervention.

As we can see, the failure of central planning as an innovation system, and the success of capitalism in many countries in supporting high levels of invention and diffusion, does not mean that the best thing a government can do to support innovation is to get out of the way. The essential role of government in successful innovation systems—military contracts in Silicon Valley and worker training in Germany, for example—suggests quite the opposite.

To understand why, in the next three sections we explore three aspects of invention and diffusion that make the innovation process a challenge to public policy, and why it is so difficult for other countries to copy Silicon Valley or German innovation systems.

These are:

- *External effects and the problem of coordination among innovators:* A firm's successful invention almost always has either positive or negative effects on the value of other firms' investments in the innovation process. Owners of a firm who are concerned solely about their profits will fail to take into account these external effects.
- *Public goods:* Innovation can be seen as the production of new knowledge by the use of a combination of old knowledge and creativity. The fact that most forms of knowledge are non-rival—making it available to an additional user does not mean that some current user will be deprived of its use—makes the innovation process one that uses public goods to produce other public goods.
- *Economies of scale and winner-take-all competition:* Big is beautiful when it comes to the knowledge-based economy: average costs fall as more units of a good or service are provided, and this means that firms entering a market first often can take the entire market, at least temporarily.

Recall from Unit 10 that these three characteristics are all among the sources of market failure introduced there; simply letting market competition regulate the process of innovation will not generally result in an efficient outcome. These same three aspects of the innovation process also pose challenges to governments that seek to address these market failures. This is because governments typically lack the necessary information (or the motivation) to develop appropriate policies.

We begin with a model of the problem of external effects and the problem of coordination among innovators, simplified to just two firms considering investing in innovations, and a government that may intervene to assist in the innovation process.

20.3 EXTERNAL EFFECTS: COMPLEMENTS, SUBSTITUTES, AND THE PROBLEM OF COORDINATION

Innovations considered by a firm typically will either increase or decrease other firms' profit levels, and affect those firms' choices about innovation. Think about just two firms, each considering innovations which are either:

- *Complements:* The value of one innovation depends on the presence of the other. The internet and the many new applications for internet use are an example; the internet would not have been worth much without the applications, and vice versa. Tin cans were invented to store food in 1810 by Peter Durand, a British merchant, and the first canning factory began production in 1813. But the cans were very difficult to open and not widely used until 1858, when Ezra Warner invented a simple can opener.
- *Substitutes:* The two innovations are valuable alone, but less valuable when some other innovation has already occurred. Either Sony Betamax (1975) or JVC's rival VHS (1976) would have been a perfectly good format for home video recording, had only one become established initially. But the introduction of the other did not add much value when one already existed in a market.

In the absence of explicit government policies or private means of coordination among firms, the challenges posed by complementary innovations and substitute innovations are quite different:

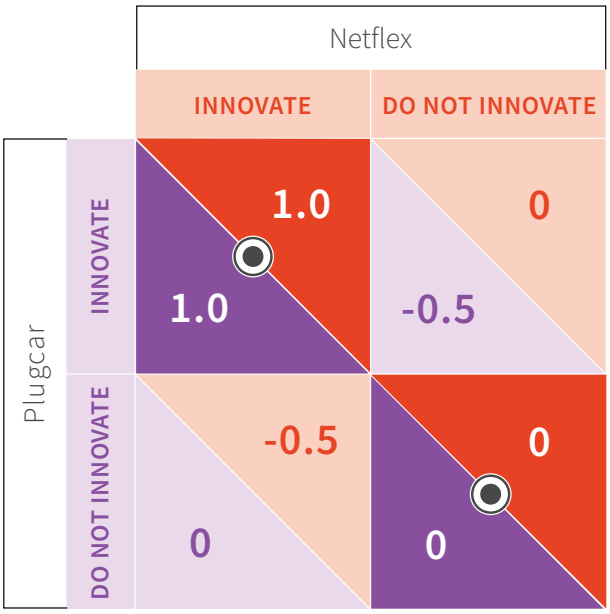
- *When potential innovations are complements:* Innovations sometimes do not occur when it would have been socially beneficial, and profitable to the firms, if they had occurred.
- *When potential innovations are substitutes:* Both innovations sometimes occur, when either one or the other would have been both more socially beneficial and more profitable to the firms involved. Competition between substitutes may impose a high cost on both innovators, as the video format war between Betamax and VHS in the 1980s demonstrated.

We can use game theory to understand how two potential innovating firms interact strategically, and show why these contrasting problems arise and why they may be difficult to solve. (You may wish to review the introduction to game theory in Unit 4, and the examples of its use in Units 12, 13, 17 and 18.)

Innovations that are complements

Here we have two hypothetical firms, Plugcar, which is considering developing a novel electric car, and Netflex, which is weighing up the likely profits and costs of investing in a mobile network of battery exchanges. As above, the presence of

Netflex makes Plugcar more valuable and vice versa, so they are complements. They will make their decisions (*Innovate, Do not innovate*) independently, but they know the profits and losses that will result in each of the four possible outcomes. They are given in the payoff matrix below. The row player is Plugcar, and its payoffs come first in each cell; the column player is Netflex, its payoffs are second in each cell. Positive numbers are profits for the company, negative numbers are losses.



Finding the Nash equilibria

Begin with the row player and ask: "What would be the best response to the column player's decision to innovate?" The best response would be *Innovate*, since the payoff is 1 rather than 0. Place a dot in the top left-hand cell. Then ask what the row player's best response would be to the column player's choice of *Do not innovate*: the answer is *Do not innovate*. Place a dot in the bottom right-hand cell. Now turn to the column player. What would be the best response to the row player's strategy of *Innovate*? The answer is *Innovate*. Place an open circle in the top left-hand cell—there will now be a dot inside a circle. Do the same for the column player's response to row player's strategy of *Do not innovate*. There is now a dot inside a circle. Wherever there is a dot inside a circle in a cell, this is a Nash equilibrium because it shows that each player is playing the best response to what the other does.

Figure 20.5 *The decision to innovate when products are complements.*

Imagine that you are Plugcar. If you do not innovate you will get zero, whatever Netflex does. If you knew that Netflex was not going to introduce its product, then you surely would not develop the Plugcar. What if Netflex does introduce its product? If you innovate you will get profits of 1. But you also stand to incur losses of 0.5 if Netflex does not innovate.

Unless you are pretty sure that Netflex is going to innovate, you may decide that you have better uses for your funds. If Netflex reasoned the same way, then neither firm might innovate even though had one done so they would have profited (not to mention the users).

DISCUSS 20.2: COMPLEMENTS

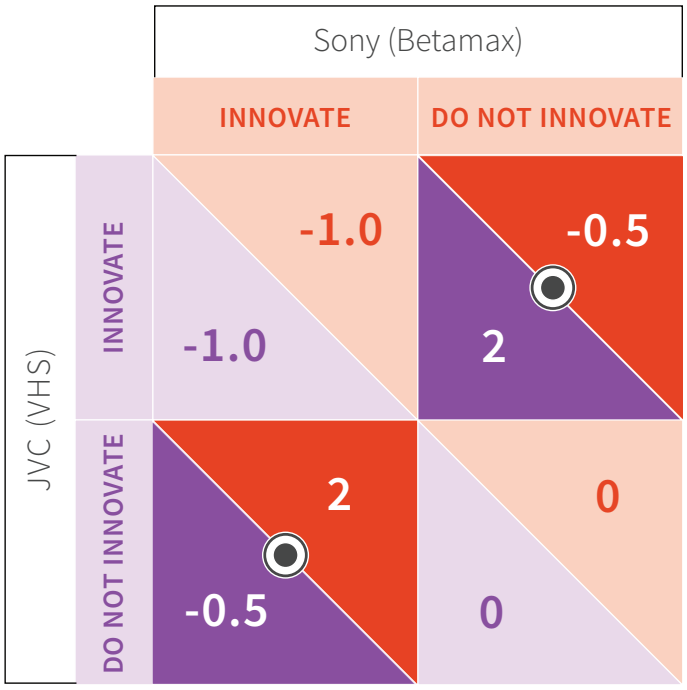
1. List some pairs of innovations that are complements, and some that are substitutes.
2. In the game in Figure 20.15, how sure would each firm have to be that the other would innovate in order to make innovating a good decision, in terms of the company's profits? Explain your answer.

Innovations that are substitutes

When two innovations are substitutes we have the opposite problem. A good example is the video format war during the 1980s between two competing standards, VHS (for "video home system" developed by Victor Company of Japan, called JVC) and Sony's Betamax format. Material using one format could not be played on machines designed to play the other, so both companies had an interest in their format becoming the most widely accepted. In 1969 Sony had collaborated with JVC and another company called Matsushita to produce a common format, but then broke away to produce Betamax.

We consider two hypothetical firms based on the Sony-JVC case. Here is the payoff matrix facing them. JVC is the row player, and Sony is the column player. As before, the first entry in each cell is the payoff of the row player.

If Sony is sure that JVC will innovate, then it will face a costly battle with big losses if JVC wins. The payoffs in the upper left-hand cell are negative for both firms because the costs of developing the new product, and competing for market share, do not offset the uncertain prospect of profits should they win. Of course, if Sony knew that JVC was not going to invest, or if it was sure it would win a not-very-costly battle with its product should both invest, then Sony would definitely invest and enjoy the winner-take-all profits, while inflicting losses on JVC.



Finding the Nash equilibria

Begin with the row player and ask: "What would be the best response to the column player's decision to innovate?" The best response would be *Do not innovate*, since the payoff is -0.5 rather than -1.0. Place a dot in the bottom left-hand cell. Then ask what the row player's best response would be to the column player's choice of *Do not innovate*: the answer is *Innovate*. Place a dot in the top right-hand cell. Now turn to the column player. What would be the best response to the row player's strategy of *Innovate*? The answer is *Do not innovate*. Place an open circle in the bottom left-hand cell—there will now be a dot inside a circle. Do the same for the column player's response to row player's strategy of *Do not innovate*. There is now a dot inside a circle. Wherever there is a dot inside a circle in a cell, this is a Nash equilibrium because it shows that each player is playing the best response to what the other does.

Figure 20.6 *The decision to innovate when products are substitutes.*

DISCUSS 20.3: SUBSTITUTES... AND COMPLEMENTS

1. Go back to Figure 4.15 and consider the game between Bettina and Astrid, in which they choose whether to use two different programming languages, C++ and Java. Describe how this game is similar to, or how it differs from the Sony-JVC game depicted here.

2. You wish to make innovating a good decision for your company's profits. How sure would you have to be that the other firm would not innovate?

Consider now that decisions in Figures 20.5 and 20.6 are made sequentially rather than simultaneously. In the case of substitutes (Sony and JVC), imagine that JVC developed its product and put it on the market (or at least convinced Sony that it would definitely do this). In the case of complements (Plugcar and Netflex), assume that Plugcar could convince Netflex that it will definitely bring the new electric car to the market.

3. Explain what the outcome in those cases would be if the two firms made their decisions sequentially rather than simultaneously.

The result is that there is sometimes too little innovation for the good of society when ideas are complementary, and too much when the innovations are substitutes.

The role of public policy

Complements

If the payoffs in the matrix were known to everyone, then a wise government would know that the top left (*Innovate, Innovate*) in Figure 20.5 is the best outcome for society. It could, in the case of complementary innovations, simply provide both firms with sufficient subsidies that both would find it profitable to make the investment even were the other to fail to do so. Or, more reasonably, it could help the two firms to cooperate in the innovation process, promising not to prosecute them for any anticompetitive practices if coordinated decision-making is prohibited by antitrust or other law.

But public policy to avoid an unfavourable outcome is a greater challenge than our simple model would suggest. There are likely to be more than two potential innovators; many proposed designs for electric cars and for recharging systems. The government would have to choose the cooperating firms, and the terms under which the cooperation would occur. In this case, companies have incentives to spend resources to influence government decisions (lobbying). As we shall see in Unit 21, there are many reasons why governments may fail to achieve the socially beneficial outcome in cases like this.

Private exchanges might have a role to play here. If the firms themselves have better information than the government, they might engage in private agreements. This is the equivalent to the bargaining among private economic bodies that occurred in Unit 10 to provide an alternative to government regulation of the use of chemical weedkillers.

Finally, firms with promising complementary innovations might agree to merge so that, as a single company, the problem of coordinating their innovation decisions would be internal to the firm.

Substitutes and standards

The substitutes in Figure 20.6 presents similar challenges for government policy. There may be a great many competing substitute innovations. Sony's Betamax and JVC's VHS were not the only entrants in the early stages of the formatting wars. Governments may lack the relevant information, or may be under the influence by one of the contestants.

Sometimes—as we will see later—one competitor's technology wins over the other: eventually, Betamax died out and VHS became the universal home videotape standard. Sometimes companies in an industry apply the same standards, because consistency increases the size of the market, and so all benefit. An example is the way the shipping industry implemented the standard for the size of containers they carry, which allowed trucks and ports to become more efficient, and therefore achieve economies of scale.

Often, however, public sector agencies play an important role in encouraging agreement among all the firms in an industry about technical standards. These are usually international bodies, like the International Telecommunications Union or the European Commission. The EU, for example, helped mobile phone companies to agree on the GSM standard for phone handsets and networks, which enabled all the manufacturers and operators to benefit from a rapidly growing European mobile market, and enabled consumers to benefit from the ease of calling other networks and declining prices.

20.4 ECONOMIES OF SCALE AND WINNER-TAKE-ALL COMPETITION

Innovation involves developing new knowledge, and putting it to use. Recall that economically, knowledge is unusual in two ways: it is a public good (what one consumes does not subtract from what is available to others) and its production and use are characterised by extraordinary increasing returns to scale. We discussed information as a public good in Unit 10. In this section we discuss the two ways in which knowledge-intensive innovation creates economies of scale.

The supply side: First copy costs and economies of scale in production

The first copy of new knowledge is costly to produce, but virtually costless to make available to others. Because *first copy costs* are large relative to the costs of making additional goods available (variable costs, or marginal costs), information production and distribution is different to any other part of the economy.

Michael Jackson's *Thriller* is the best-selling music album in history. It cost $750,000 to produce in 1982 (about twice that amount in 2015 dollars). The marginal cost of producing additional copies is less than $1 for a CD, and almost nothing if it is a download. A CD sells for about $10, and a download for the same amount. The first copy cost of even a modest production by a new band will be at least $10,000, with about $1 marginal cost for each CD.

To develop a new high-quality textbook in the US costs between $1m and $2m, to compensate the writers, designers, editors and others for their work. This is the first copy cost. The cost of producing and distributing the physical books (printing, warehousing, and delivery included) for a successful text are approximately $12 per book. This is its marginal cost. American students know that year-long introductory textbooks typically sell for ten times this amount.

An international team of authors volunteered their expertise without compensation to write this introduction to economics, but its first copy cost is the design and production of the interactive ebook. Because it is available electronically, its marginal cost is zero.

New software is another example. According to John Miller, who for 25 years was a software engineer at Microsoft, Amazon, and other tech companies, the cost of developing the first copy of a software package like Microsoft Windows, if you were starting from scratch today, would be approximately $18.75bn. The marginal cost of making additional copies available would simply be the cost of the time taken to install it on a computer.

The first copy of movies and computer games may be extraordinarily expensive. The production budget for *Star Wars: The Force Awakens* (released in 2015) was $200m. The development cost for the computer game *Star Wars: The Old Republic* (2011) was between $150m and $200m. These figures do not include the marketing and promotion costs, such as advertising, that should be included in the first copy cost, and may be bigger than the production costs. Now that movies are distributed digitally to cinemas, making a film available costs virtually nothing. The marginal costs for movies or games sold on DVD are around the same as for a CD, and when they are sold as digital downloads, they are zero.

Some of the most striking examples of first copy costs bring us back to the HIV/AIDS treatments at the heart of the patent controversy in South Africa. The average first copy cost of a new drug according to a study in the US in 2003 was $403m. This fact

explains the difference in price between drugs that are still under patent, giving the producer a temporary monopoly, and the prices that users pay once the patent has expired so that other producers compete with the originator of the drug.

Here is an example. Omeprazole, a very widely prescribed dyspepsia drug, was patented and launched in 1989 by the firm Astra Zeneca (then called Astra AB) and sold across Europe and the US (where it was sold under the brand name Prilosec). In the US the patent expired in 2001, and by 2003 28 tablets of brand-name Prilosec sold for $124, while the equivalent packet of generic Omeprazole cost only $24. Today the generic form of the drug is manufactured by more than 30 companies, with prices in the US as low as 16c per tablet.

A wide range of knowledge-intensive goods or services, including a new car and aircraft designs, show similar differences between first and subsequent copy costs.

In Unit 7 we studied how a firm sets prices, and how it decides how much to produce. In Figure 20.7 we show a set of cost curves for a firm producing a knowledge-intensive good. The numbers are hypothetical, and they understate the true size of the first copy cost relative to marginal cost. Even so, the vertical axis is not drawn to scale so we can read the figure.

- *Total cost:* The curve starts at the first copy cost, and then rises very little with increased production.
- *Marginal cost:* The curve is low and constant.
- *Average cost:* The curve (including economic profits and the first copy costs) falls as quantity increases, as the cost of the first copy is spread over larger units of output.
- $MC < AC$: No matter how many units are produced, the marginal cost will always be less than the average.

A firm producing a knowledge-intensive good that wants to make economic profits will have to cover its first copy cost. Therefore it must charge a price greater than the marginal cost: the price will have to be at least as high as the average cost curve.

This means the production of knowledge-intensive goods cannot be described by the competitive markets of Unit 8 in which price equals marginal cost ($P = MC$), but instead by the model of price-setting firms in Unit 7. In Unit 7 we assumed that $P > MC$ because of limited competition. Here is an unavoidable consequence of first copy costs, and no matter how many competitors there are, price cannot be competed all the way down to marginal cost.

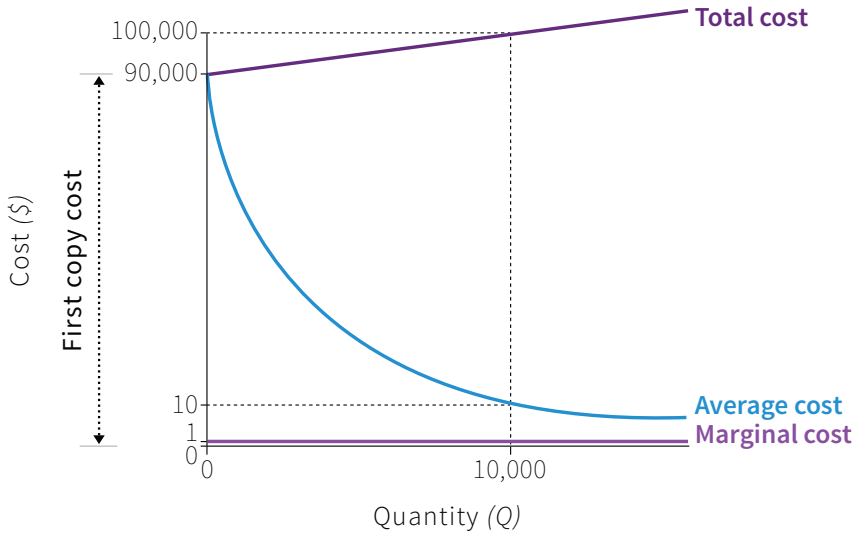


Figure 20.7 *The first copy, marginal, and average costs of a knowledge-intensive good.*

The demand side: Economies of scale through network effects

Also, the value of many forms of knowledge increases when more people use it. Because the benefits to users increase as the network of users grows, demand-side increasing returns are sometimes called *network external effects*. The external effect is that when one more person joins the network, all others benefit.

Languages are a good example. Today, more than 1 billion people are learning English, more than three times as many people as speak English as their first language. The demand for English does not derive from the intrinsic superiority of the language, or because it is easy to learn (as many of you will know), but simply because so many other people, in many parts of the world, speak it. There are many more people who speak Mandarin (Chinese) and Spanish as a first language, and almost as many Hindi and Arabic speakers, but none of these languages is as useful to communicate globally as is English.

Having a particular games console is better when lots of people have the same one, because developers will produce more games for it. A credit card is more useful when many people have the same card, because lots of shops will accept it as payment.

But have you ever wondered who bought the first telephone, and what they intended to do with it? Or what you could do with the first fax machine?

The technology behind the fax, a device to send images of documents over a telephone line, was first patented by Alexander Bain in 1843—although his image-sending innovation had to use the telegraph, because nobody had invented a telephone yet. A commercial service which could transmit handwritten signatures using the telegraph was available in the 1860s. But the fax remained a niche product until 120 years later when it became so popular that, in less than 10 years, almost every office installed its own fax machine.

This tells us the first thing we need to know about demand-side economies of scale: there is little incentive to be the first to adopt a technology with this characteristic.

The second thing we need to know is that, if two versions of this type of technology are competing, the one that gains a larger number of adopters at the outset will have an advantage, even if the other one is cheaper or better. To see this, let's take another look at the video format war between Sony and JVC.

Sony's Betamax format was superior to JVC's VHS for its picture and sound quality. But in the early 1980s Sony made a strategic error by limiting the record time to 60 minutes. If customers wanted to use their new Sony Betamax to record a feature film, they needed to change the tape in the middle of the recording. By the time Sony had extended its recording length to 120 minutes, there were so many more VHS users that the Betamax format all but disappeared.

The video formatting war, and its outcome, is an example of *winner-take-all competition*, in which economies of scale in production or distribution give the firm with the largest share of the market a commanding competitive edge. Winner-take-all competition does not necessarily select the best.

To see how this works, Figure 20.8 depicts competition based on the Sony and JVC case. The length of the horizontal axis is the number of people purchasing either Sony's Betamax or JVC's VHS. We assume that the price of the two products is identical. The net value to any consumer of purchasing Sony Betamax depends on the quality of the product, q^S (the superscript "S" is for Sony), how many others are using the product, n^S, minus the price paid, p. The number buying Betamax is measured from the left to the right, starting at zero and extending potentially all the way to the entire market. Thus the net value of benefit to a consumer of Betamax is given by the rising blue line (its equation is just $\pi^S = q^S n^S - p$). If everyone buys Betamax, the value to each purchaser is shown in the figure, π^{Smax}, which is equal to $q^S n^{total} - p$.

In the same figure the net value of JVC's product VHS is given by the red line whose equation is $\pi^J = q^J n^J - p$ (as before, "J" stands for JVC). Because there are only two firms competing, the number buying JVC VHS is just the total, minus the number buying Sony Betamax.

Let's assume that the Sony format is better, in the sense that if everyone bought it the net value would be greater than if everyone bought JVC's format, that is $\pi^{Smax} > \pi^{Jmax}$. This is illustrated in the figure by the fact that the height of the blue Betamax line where it hits the right-hand axis (everyone using Betamax) is above the intercept of the red VHS line with the left-hand axis (everyone using VHS).

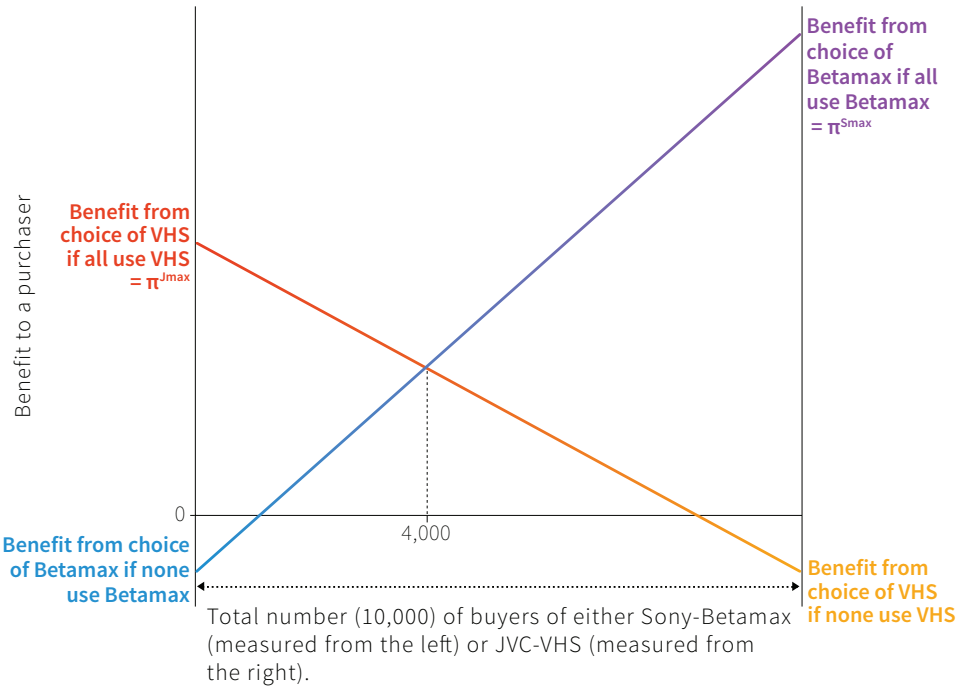


Figure 20.8 *The value (net of costs) of becoming part of a network.*

The first thing to notice is that if at a particular moment everyone is buying VHS, then a new buyer will certainly prefer VHS to Betamax. To see this in the diagram, look at the left-hand side and consider a new buyer. For this person, the value of VHS is high (the intercept with the left-hand axis); the value of Betamax is negative. This is because the new user would have to pay the price of the Sony product, but would not get any benefits because there are no other users. This is true even though we have assumed that the Sony product costs the same as the JVC product, and that Sony has the better product. In the video format example, this would be a situation in which Sony makes a higher quality cassette but there is no content.

The second lesson from the figure is that even if many consumers (but fewer than 4,000) were buying the Sony product, the new consumer would still prefer VHS. For Sony to break JVC's monopoly, it would have to get at least 4,000 buyers to purchase its product. Then Sony rather than JVC would be the winner, and eventually take all.

So winner-take-all is not actually a competition, and the winner need not be the better alternative. This is sometimes called *lock-in*.

But this is not the whole story. The history of innovation in the knowledge economy is full of more complicated stories, in which changes are constantly occurring, for many reasons. For example:

- *Browser wars:* When the internet became popular, the market for internet browsers was dominated by a product called Netscape Navigator. It was displaced by Microsoft Internet Explorer in the "browser wars" of the early 2000s. Internet Explorer, in turn, was later challenged by Mozilla Firefox and Google Chrome.

- *Smartphones:* At the beginning of 2009, Android smartphones had a market share of 1.6%, Apple's iPhones had 10.5%, and the market was dominated by a technology called Symbian, with 48.8% share. At the beginning of 2016, 84.1% of smartphones sold were based on Android, Apple's smartphones had a share of 14.8%, and Symbian smartphones were no longer being manufactured.
- *Social networks:* In June 2006, 80% of people who used a social network used a site called MySpace. By May 2009, more people used Facebook than MySpace.

DISCUSS 20.4: THOMAS JEFFERSON

Thomas Jefferson (1743-1826), America's third president, noted the peculiar and wonderful nature of an idea when he says that:

"Its peculiar character… is that no one possesses the less, because every other possesses the whole of it. He who receives an idea from me, receives instruction himself without lessening mine; as he who lights his taper [candle] at mine, receives light without darkening me."
Thomas Jefferson to Isaac McPherson, *Writings* (1813)

Jefferson went on to say something that even then was controversial:

"It would be curious, then, if an idea, the fugitive fermentation of an individual brain could… be claimed in exclusive and stable property."

To him, granting to an individual the exclusive right to own and exclude others from the use of an idea just did not make sense, any more than it would make sense for a person to refuse to tell someone what time of day it was.

1. Rewrite the first part of Jefferson's quote using the economic terms you learned in this course.
2. Why would Jefferson's the second part of Jefferson's quote be controversial, even before the advent of modern knowledge intensive goods like new automobiles or aircraft?

20.5 MATCHING (TWO-SIDED) MARKETS

A market is a way of putting together people who might benefit from exchanging a good or service. Often these are potential buyers and sellers of the same commodity, such as milk, and the sides of this market are farmers supplying milk and consumers demanding it. In common usage a market may also be a place such as the Fulton Fish Market that we described in Unit 8, or a place where those selling fresh vegetables, cheese and baked goods congregate, knowing that they will encounter potential customers. In these markets buyers do not care about who produced the fish or the milk that they buy; and sellers similarly are not concerned about who is buying, as long as they buy.

Matching (two-sided) markets

But people also use the term *market* to describe a different kind of connection, in which the people on each side of the market care who they are matched with on the other side. This is what people have in mind when they speak about the "marriage market", for example. Most of us do not get married in the way that we get a carton of milk in the grocery market. The marriage market is about getting married to a person with the combination of characteristics that you find most desirable in a spouse. We call markets like these *matching*, or *two-sided, markets*.

MATCHING MARKETS

A matching market is also called a *two-sided* market because:

- It matches members of two distinct groups of people
- Each person in the market would generally benefit from being connected to the right member of the other group

In our video Alvin Roth, an economist who specialises in how markets are designed (and who won the Nobel prize for his work on the subject in 2012), explains how matching markets function.

We have recently seen a proliferation of online platforms that connect individuals in two groups, starting with the launch of consumer-to-consumer trader eBay in 1995. These platforms make up a general-purpose technology that allows the participants to benefit from being networked together, and so are examples of two-sided markets (you can also describe them as two-sided networks).

An example is Airbnb, a service that connects travellers looking for short-term apartment rentals with owners seeking to make money by making their home available while they are not living in it. Airbnb is a platform that puts the group of apartment-seekers in touch with the group of apartment owners who would like to

offer their apartments for rent. Tinder does the same thing for people who want to find a date for the evening. A service called JOE Network puts employers in contact with people who have recently been awarded PhDs in economics.

These matching platforms have become important in economics because of the magnitude of the network connections that are now possible. But while connections on this scale are now technically feasible, there is no mechanism that will reliably bring two-sided markets into existence even if they would create gains for the participants on both sides.

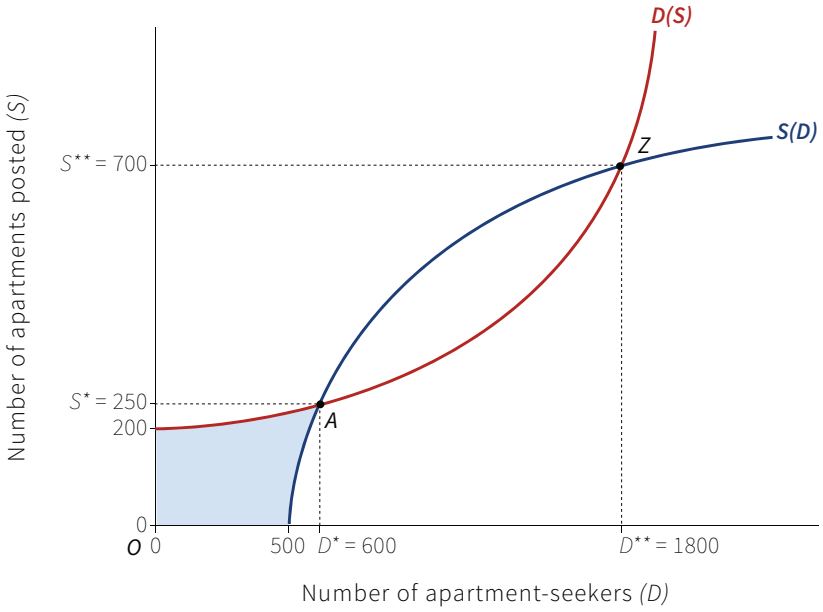
At an early stage, these markets—meaning the creation of the platform, or the marketplace, or whatever it is that connects people—face a chicken-and-egg problem. Think about Airbnb: it makes money by charging a commission on each deal that is struck. Unless there are a large number of apartment-seekers consulting its web site, there is no reason for an apartment owner seeking a rental to offer an apartment for rent. Without apartments to rent, Airbnb will not be able to make money, so there would be no incentive to create the platform in the first place.

A model of a two-sided matching market

In economics, these two activities—going to Airbnb's web page, and posting one's apartment on it—are termed *strategic complements*, meaning that the more of the first that occurs, the more benefit there is for doing the second; and the more benefit there is for doing the first, the more of the second that occurs. This can also be described as demand-side increasing returns to scale (or a network externality) that, as we have seen, are characteristic of knowledge as an economic entity.

The terms complements and substitutes can be applied both to things and to strategies. We talk of strategic complements when the strategy to introduce one innovation enhances the benefits from the introduction of another innovation, and *strategic substitutes* when the strategy to introduce an innovation has the opposite effect.

Figure 20.9a illustrates the chicken-and-egg problem. We begin with the demand side: the apartment-seekers. As more apartment-seekers are looking at the site (moving to the right along the horizontal axis), then more suppliers will post their information. The demand for apartments, D, depends on the supply of them, S. As long as a minimum number of apartments are posted on the site (200), then some people will look for an apartment there; the more that are posted, the more people will look. The curve representing the demand for apartments as a response to the supply (the number posted) is labelled $D(S)$.



If no apartment-seekers are consulting the site

No apartment suppliers will post their information. Nobody doing anything is therefore a Nash equilibrium, as shown by *O*.

Point *A*

At A, supply and demand curves intersect. It is another Nash equilibrium.

Point *Z*

Point *Z* is also a Nash equilibrium.

Figure 20.9a *Supply and demand in a two-sided matching market: The case of Airbnb.*

We turn now to the supply of apartments on AirBnB. The minimum number of apartment-seekers on the site that will induce even a small number of suppliers to post their apartment there is 500 (look at where the supply curve intercepts the horizontal axis). As the number of apartment-seekers viewing the site rises beyond 500, an increasing number of suppliers will post their information. The supply curve is labelled *S(D)*. But there is a limit to how many people will want to rent out their home temporarily, so *S(D)* flattens out as we move to the right (it is concave).

The situation is similar for apartment demanders. As long as a minimum number of apartments are posted on the site (200), then some people will look for an apartment there, and the more that are posted, the more people will look. Note that these are not ordinary demand and supply curves: the demand by apartment-seekers depends on the supply of apartments, which is why this is written as *D(S)* in Figure 20.9a. Similarly, the supply of apartments depends on the number of those searching and is written as *S(D)*.

Now consider the problem if you have created Airbnb, and want people to start using the site regularly. Suppose there are 220 suppliers posting their information and 550 demanders looking. What will happen? If this information is known to both suppliers and demanders, then we can see from the figure that the number of suppliers is in excess of demand; given the weak demand, suppliers will withdraw from the site. As they do so, the number of apartment-seekers will fall. The downward spiral continues until nobody uses the site. In this case the costs incurred by Airbnb (writing the software, maintaining the platform) will not be compensated.

From this, we learn two things:

- *There is a Nash Equilibrium at which there is no Airbnb:* At point *O*.
- *The Nash Equilibrium at point A is not stable:* At *A* the supply and demand curves intersect, with 250 apartments posted and 600 people seeking apartments. If just one apartment or apartment-seeker drops out, we will be in the blue zone; and if we enter the blue zone, we know that the market will collapse to *O*.

If, however, there is a sufficient number of demanders (greater than 600), and the number of suppliers is greater than 250 (D^* and S^* in Figure 20.9a), then the number of both will grow until there are 700 suppliers and 1,800 demanders (Figure 20.9b). Once supply is greater than 250, demand is above supply (*B*), which encourages new suppliers to list their site (*C*), which in turn attracts new apartment-seekers. In this case, there is an upward spiral until point *Z* is reached.

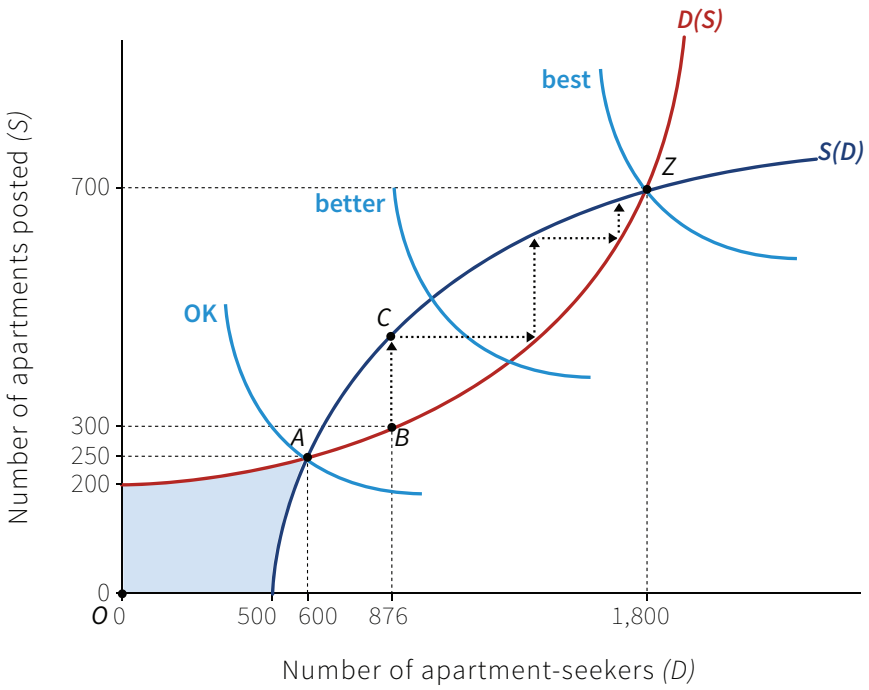


Figure 20.9b *Supply and demand in a two-sided matching market: The case of Airbnb.*

Like the situation of zero activity (*O*), *Z* is a stable Nash equilibrium: given that there are 700 suppliers posting their apartments there are 1,800 demanders looking, and conversely given the number of lookers there will be 700 suppliers. To check that it is stable, imagine a situation with one more, or one less, apartment-seeker, or one more, or one less, supplier.

So we could be in a situation with no Airbnb or, once there is enough balance between suppliers and demanders, a situation in which the popularity of Airbnb grows until there are a large number of people using the site. To see that the second outcome is preferable, we have drawn indifference curves representing the total social benefits created by each of the technically feasible combinations of suppliers and demanders on the site in Figure 20.9b.

The total benefits (labelled "OK", "better" and "best") include the gains to the suppliers (the rent they receive) and to the demanders (value of the accommodation they enjoy, minus the rent they pay). The curves have the familiar shape: more of each creates greater total benefits, but expanding both suppliers and demanders at the same time is more advantageous than increasing one without increasing the other. This is just a result of the fact that the two activities—searching and posting on either side of the platform—are strategic complements.

Market failures in matching markets

The economic policy challenge is to find a way to ensure that someone will create the platforms that create benefits for participants that are sufficient to justify the cost. This is sometimes done by the public sector playing a role in creating the platform, as it did in the case of the internet, or physical marketplaces in cities and towns. But in many cases (such as Airbnb, Tinder and many other private platforms), the existence of a two-sided market is the haphazard result of a forward-looking individual having both the idea and the resources to launch a large, risky project.

One strategy for solving the chicken-and-egg problem is for companies to charge low or zero prices to one group of users, which then attracts the other group. For example, to read this ebook as a portable document format (PDF) file, Adobe lets you download its PDF reader for no cost. If many people read documents as PDFs, it incentivises document creators to pay for Adobe Acrobat, the software used to create PDF files.

While some two-sided markets, such as Wikipedia, are not designed to be money-makers, most are. And some of those who succeeded in creating widely-used platforms have gained extraordinary wealth. In 2015 Facebook was valued at $212bn and Mark Zuckerberg, who founded the company, owns a little over a quarter of it.

These innovation rents, unlike those associated with a new technical innovation like the spinning jenny studied in Unit 2, may not be competed away because would-be competitors face the very same chicken-and-egg problem that the successful innovators solved.

The problem is similar to the example of the strategic interaction between Plugcar and Netflex discussed earlier in this unit. There are probably many potentially mutually beneficial two-sided markets that do not exist (or do not exist yet) because of the chicken-and-egg problem. For instance, there has been little new competition in the credit card industry: it would be difficult to persuade merchants to accept a new type of card if not many shoppers carried it, and it would be difficult to encourage shoppers to carry a card that not many merchants would accept.

DISCUSS 20.5: CHICKEN-AND-EGG

Platforms such as Airbnb, Uber, YouTube, eBay, Twitter, Facebook and Wikipedia have successfully overcome the chicken-and-egg problem mentioned above.

1. What are the gains these platforms offer, and which other markets have they disrupted?
2. What factors made it possible for these platforms to disrupt existing markets?

A catalogue of policies

The last three sections have introduced three reasons—external effects, public goods and economies of scale—why market competition for profits cannot create an efficient innovation process by itself. Public policies can encourage useful innovations and accelerate their diffusion to all users who may benefit. We have already mentioned the need for complementary infrastructure and the possible coordinating role of government-set standards.

In the next three sections we study two of these policies:

- *Intellectual property rights:* These policies support innovation rents accruing to successful innovators.
- *Subsidising the supply of inputs to the innovation and diffusion process:* These policies provide and subsidise basic research, education and prizes for successful innovations that are then placed in the public domain, and ensure the low-cost dissemination of information.

20.6 INTELLECTUAL PROPERTY RIGHTS

Patent protection may be unnecessary for an innovator if secrecy is possible, or social norms prevent copying. The formula for Coca-Cola has famously remained a secret for 100 years. The company claims it is known by only two executives at any time, who never travel on the same aeroplane. A chef's signature dish is not a secret, but social norms among chefs would make the costs of copying a recipe without permission extraordinarily high. Comedians rarely steal each other's jokes for the same reason.

In other cases, an innovation may be known, but barriers to copying can be built into the product itself. Digital watermarking technology allowed some music distributors (briefly) to make recorded music that could not be copied. Seed companies successfully accomplished the same thing by introducing hybrid corn and other varieties that do not reproduce well.

Firms can also rely on superior capabilities that are complementary to a technological product to protect their innovation rents. Such capabilities could be a superior sales force, the ability to bring products to market more quickly, or exclusive contracts with input suppliers.

Secrecy, barriers to copying or complementary capabilities may not be effective against rivals who manage to invent the same product independently, or who reverse-engineer it by starting with the finished product and working out how it was made.

Where a novel idea is both codifiable (it can be written down) and non-excludable (imitation cannot be prevented), governments have created laws protecting intellectual property rights. There are very many kinds of intellectual property, but the most commonly used are patents, trademarks and copyright.

INTELLECTUAL PROPERTY PROTECTION

Codifiable and non-excludable ideas can be protected in the following ways:

- *Patents* require the innovator to disclose their idea in a patent application, which is examined by a patent office and subsequently published. If the examiners are convinced the idea is sufficiently new and inventive they will grant the innovator a patent. In most cases, a patent gives the innovator the right to take any imitator to court for 20 years: this can be extended to 25 years in the case of pharmaceutical patents. Some countries vary the length of patent protection.

- *Trademarks* give the owner of a logo, a name or a registered design the right to exclude others from using it to identify their products. Trademarks can be extended indefinitely. Patents and trademarks are generally registered at a dedicated office.
- *Copyright* gives the author of an intellectual work such as a book, an opera, or software code the right to exclude others from reproducing, adapting and selling it. Copyright is generally not registered. The author must make a claim if he or she believes it has been violated. Copyright terms are far longer than those for patents, and have been progressively extended. Copyright applies for a minimum of 25 years and in the US currently for 70 years after the death of the creator. Long copyright terms are controversial, because often the benefits go to people who did not create the work.

Historically, most people assumed that patents encouraged the development and use of innovations. Now economists and historians are taking a second look at whether intellectual property rights promote or destroy innovation. The answer depends on whether the beneficial effects of protecting innovations more than offset the impediments that these rights create for the diffusion of good new ideas and devices.

An important historical case is the steam engine, which was so important to the Industrial Revolution. There were several types of steam engine invented during the 18th century, but the most successful type was patented in 1769 by James Watt. He was an engineer, and did nothing to commercialise his innovation. In fact, he did not begin production in earnest until six years after he invented it.

The commercial value of the patent was an afterthought for Watt. The businessman Matthew Boulton bought a share in the patent, and persuaded Watt to move to Birmingham (one of the centres of the Industrial Revolution) to develop the new engine he had invented. Boulton also campaigned successfully to extend the period of the patent from 14 to 31 years.

Afterwards, Watt and Boulton used the courts vigorously to prevent any other steam engines from being sold, even if they were different to Watt's design. Among these was Jonathan Hornblower's rival invention, which was more efficient that the Watt design. Watt and Boulton challenged Hornblower's patent, eventually winning the case in 1799.

Another superior invention, created by an employee, was blocked when Watt and Boulton succeeded in broadening their patent to cover the new design, even though they had not had any part in its development. Ironically, Watt knew how to make his machine more efficient, but he couldn't make the improvement. Someone else held the patent.

Under the Watt-Boulton patent, the UK added about 1,250 horsepower of steam engines per year. In the 30 years after it expired, more than 4,000 horsepower a year of steam engines were installed in England. Fuel efficiency, which had improved little while the patent was in force, increased by a factor of five between 1810 and 1835.

When Petra Moser, an economic historian, studied the number and quality of technical inventions shown at mid-19th century technology expositions, she found that countries with patent systems were no more inventive than countries without patents. Patents did, however, affect the kinds of inventive activities in which countries excelled.

There is no doubt that patent protection is essential to the process of new knowledge creation in some industries. When the patent on a pharmaceutical blockbuster drug (a drug with annual sales of more than $1bn in the US) expires, firms specialising in copying drug formulations and selling generic versions of the drug can enter the market, and the drug's price decreases as it is exposed to price competition. The patent owner's profits decrease significantly. We say the patent owner falls off a "patent cliff". Patent cliffs demonstrate that monopolies created by patents can be immensely valuable for the patent owner, but costly for users of the patented innovation.

When the DVD was introduced, it became apparent that the technology would allow consumers to not just own, but also to copy music and films from these disks in high quality. This posed a significant dilemma for the music and film industries that was addressed through new laws making it illegal to subvert digital rights management (DRM) technology, which the film companies used to stop people copying the content without permission. These same laws are now often used when users share content that is copyright protected over the internet. Today DRM technology helps to protect the companies we now call content providers, who use the internet as a distribution device—think of a television company that streams sports events live to computers and phones.

WHEN ECONOMISTS DISAGREE

INTELLECTUAL PROPERTY RIGHTS: DYNAMO OR DRAG?

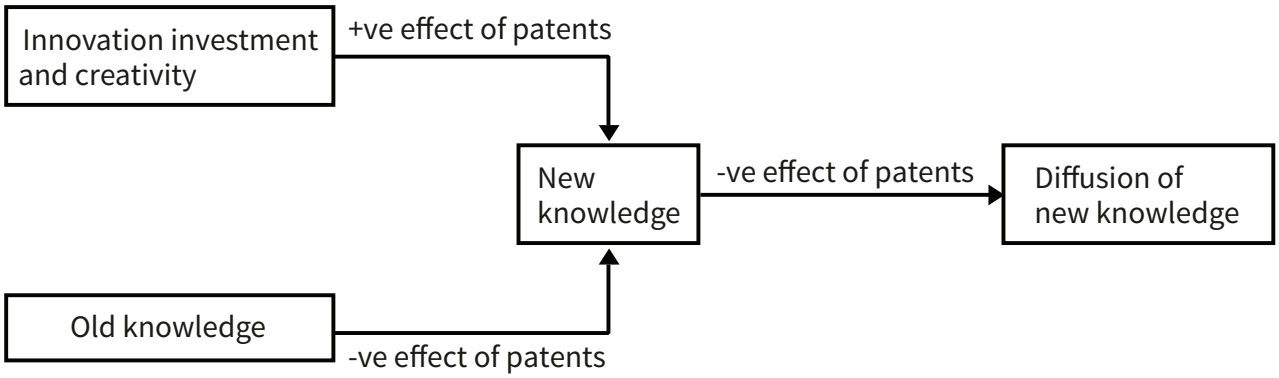
Look again at our video in which F.M. Scherer argues that patents incentivise R&D in pharmaceutical companies (unlike in many other sectors, he says), so that they continue to develop new blockbuster drugs.

In our Economist in action video, Petra Moser explains that copyright protection for 19th century Italian operas led to more and better operas being written, but she also presents historical evidence suggesting intellectual property rights may do more harm than good for the innovation process if they are too broad or too long-term.

DISCUSS 20.6: INTELLECTUAL PROPERTY RIGHTS

1. What are the main obstacles to allowing the very poor to benefit from medical innovations at marginal cost, and can these obstacles can be overcome through regulation?
2. Why does an extension of copyright terms (for example, an extension of the life of the protection) not change incentives to improve works (texts and operas) as much as the introduction of copyright itself? In your answer, consider who benefits from extended copyright terms.

If we look more closely at the innovation process as the production of knowledge by means of knowledge, we may find more consensus on the positive and negative aspects of intellectual property rights. Figure 20.10 is a schematic representation of the innovation process. Arrows represent inputs, pointing towards the aspect of innovation that they affect.



Old knowledge helps make new knowledge

Patents slow down this process. As Watt and Boulton found out, patents can impede the use of some aspects of old knowledge if it is covered by a patent.

Patents encourage innovation

The creation of new knowledge gives successful inventors recognition and innovation rents. Watt did not invent the steam engine to profit from the patent he would receive, but other innovators are strongly motivated by the prospect of commercialising their inventions.

Patents slow diffusion

Patents prevent other innovators from realising the full benefits of new knowledge after it has been created. Watt and Boulton managed to use patents to stop rival inventors from creating their own, perhaps better, steam engines.

Figure 20.10 *Patents and the production of knowledge using knowledge.*

20.7 OPTIMAL PATENTS: BALANCING THE OBJECTIVES OF INVENTION AND DIFFUSION

Patents confront us with an economic problem: how best to balance the competing objectives of making good use of existing knowledge, devoting sufficient economic resources and creativity to creating new knowledge, and diffusing the new knowledge that is created. An "optimal patent" is one that best advances the use of knowledge in the economy. Currently, agreements administered by the World

Trade Organization, which regulates international trade, may prevent countries from choosing patent length; but, given complete freedom of choice, how could a policymaker decide the optimal patent length?

In Figure 20.11, we look first at the decision of an innovator in the upper panel.

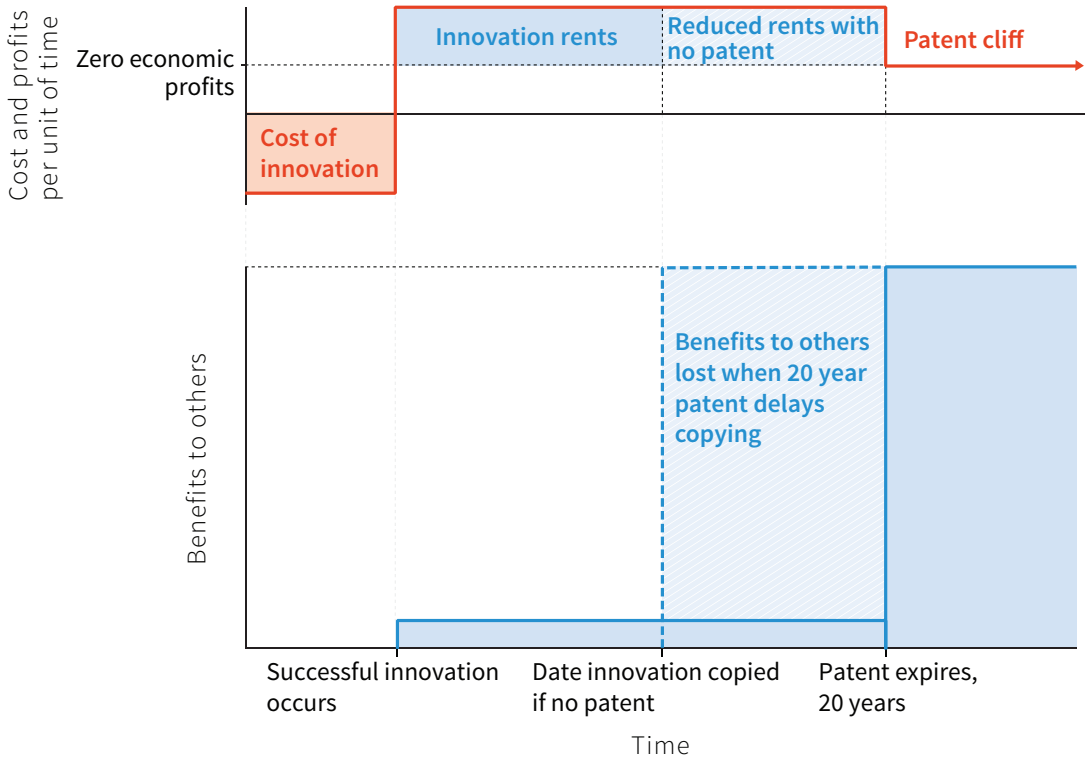


Figure 20.11 *Costs and rents associated with innovation for the inventor and others.*

The diagram highlights the timing: the innovator incurs costs in the first stage of the process, shown by the red rectangle. If the innovation is successful, the firm makes innovation rents above economic profits: this is the rectangle above the dotted line, marked as zero economic profits. The blue shading shows that, in a situation in which there is not a patent, innovation rents will disappear once the innovation is copied. Note that even without a patent in the case shown in the diagram, the innovation is not copied immediately and the innovator recovers the cost of innovation. If there is a patent, then the firm benefits from innovation rents for the life of the patent.

In the lower panel of Figure 20.11, we include the benefits to others in the economy that arise from the innovation. The first point is the obvious one that without the innovation, there are no benefits to others; the second is that the benefits to others are reduced by the duration of the patent. Earlier imitation of the innovation brings benefits to the economy shown by the dashed rectangle in the lower panel.

From this, we can see that a long patent emphasises the benefits of rapid innovation, and a short patent emphasises the benefits of rapid imitation. But we can't decide by looking at Figure 20.11 how long the optimal patent should be. To help make a decision on optimum patent length, we use a diagram that has the duration of the patent along the horizontal axis and the probability of innovation on the vertical axis.

The trade-off between the benefits of diffusion and of invention

Figure 20.12 shows how we can represent the benefits of innovation to society as a whole. On the horizontal axis are shown the total benefits to others in the economy if the firm innovates. This is called B. On the vertical axis we estimate the probability of innovation, called p. The downward-sloping curves are indifference curves called isototal benefits curves. The total benefits from innovation are:

$$\begin{aligned} \textit{total benefits} &= \textit{probability of innovation} \\ &\quad \times \textit{benefits to others if firm innovates} \\ &= pB \end{aligned}$$

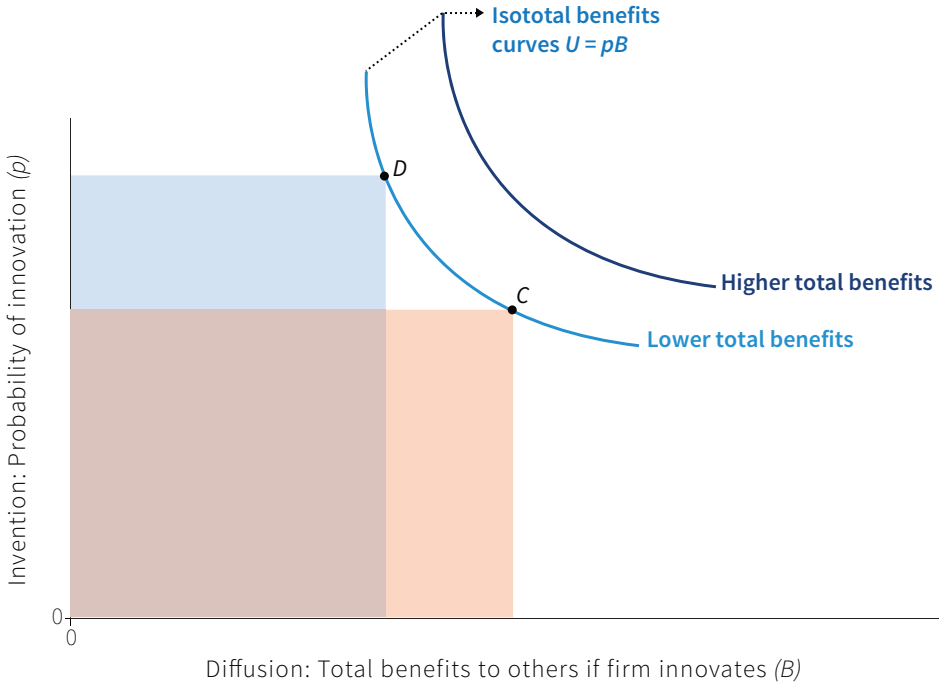


Figure 20.12 *The trade-off between the benefits of invention and diffusion: Isototal benefits curves.*

Along any one of these curves, the total benefits arising from an innovation equal to *pB* are constant (hence isototal benefits). We can see this graphically by noting that the rectangle with the corner at any point on the curve has the same area as the rectangle with the corner at any other point (since the area of each rectangle is equal to pB). Points *C* and *D* illustrate this.

In the figure there is also a preferable isototal benefits curve: every point on it has identical total benefits, and these are superior to the benefits along the inner curve. Just as in previous units, reaching the highest feasible curve is the policy objective of the government.

Feasible invention and diffusion

What are the constraints? What limits the total benefits that will occur if the innovation takes place? This will depend on the length of the patent, because a longer period of patent protection is thought at least initially to increase the probability of innovation, p, but to reduce the amount of total benefits for others, *B*, if the innovation occurs because of the delay in copying.

Even when there is no patent, innovation can occur; this is shown on the vertical axis of Figure 20.13. In these cases the innovator could capture innovation rents just by being the first, because it takes competitors a long time to catch up.

Figure 20.13 shows that as the duration of patent increases (moving to the right along the horizontal axis), so does the probability of innovation because innovation rents are protected for a longer period of time. Beyond a particular length of patent protection, however, the probability of innovation begins to decline because long-term patents will prevent other potential innovators from using protected knowledge or processes to develop an idea.

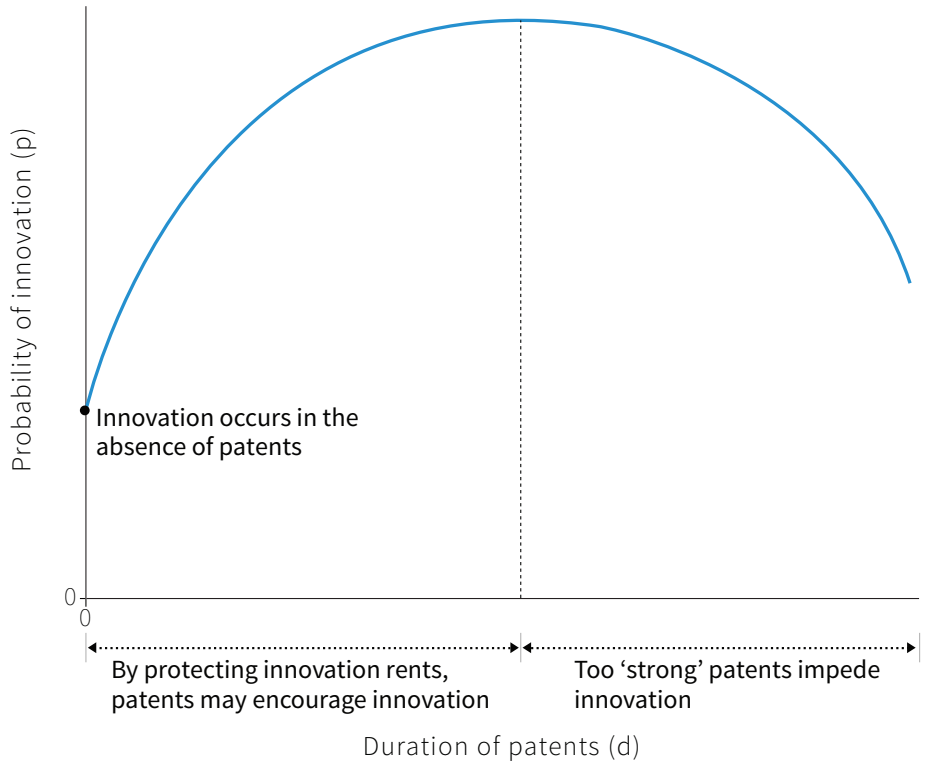


Figure 20.13 *Patent duration and probability of innovation.*

We can show the feasible set in Figure 20.14: this shows the trade-off between a higher probability of innovation and the total benefits to others if the firm innovates.

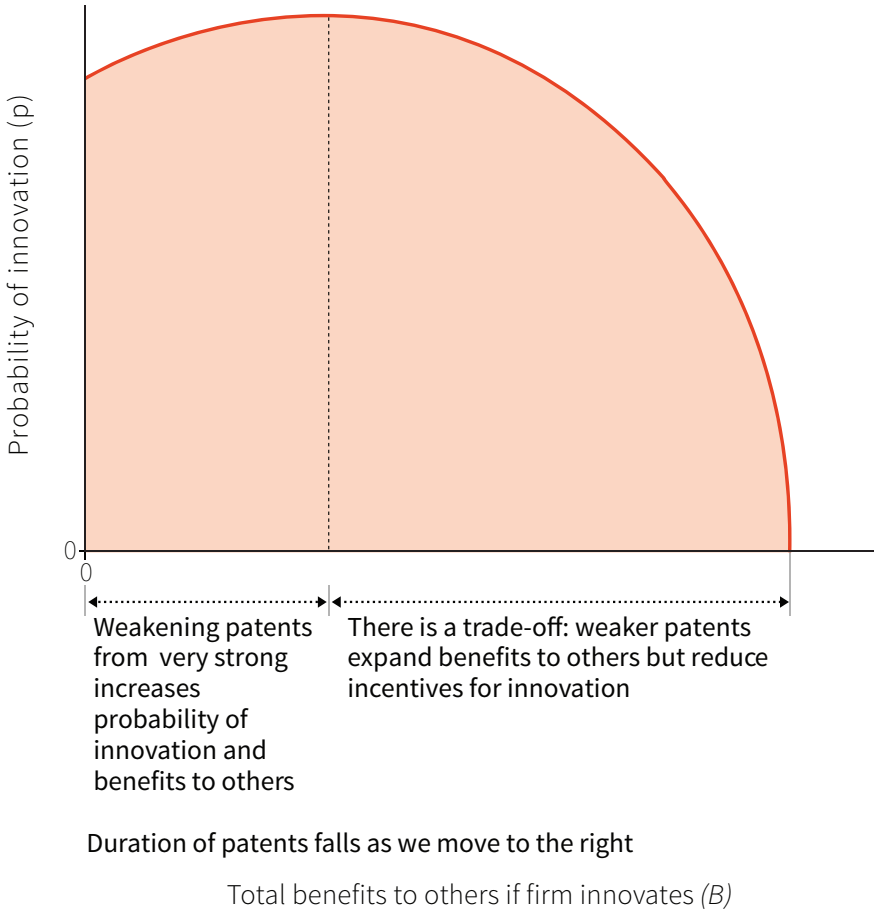


Figure 20.14 *The feasible set: Innovation probability and benefits to others.*

Each point on the boundary of the feasible set is the result of a given length of the patent, starting at the vertical axis intercept with a patent that never expires. As we move to the right, the duration of a patent declines. There are increasing benefits to others. Initially this increases both the benefits to others should the innovation occur, and (as we saw in Figure 20.13) the probability that this will be the case. This gives a positively-sloped section of the feasible set. However, as we have also seen, at some point there will be a trade-off: a further reduction in patent duration will decrease the probability of innovation, even though it expands the total benefits that would result should the innovation occur. This explains the downward-sloping portion of the frontier of the feasible set.

Optimal patent duration

If we now put the feasible set together with the isototal benefits curves, we can determine the length of the patent that maximises the expected benefits consistent with the constraints imposed by the trade-off between the incentive for innovation and stimulating diffusion. The highest attainable level of total benefits is shown by the tangency of the isototal benefits curve with the feasible set. This is point *A*.

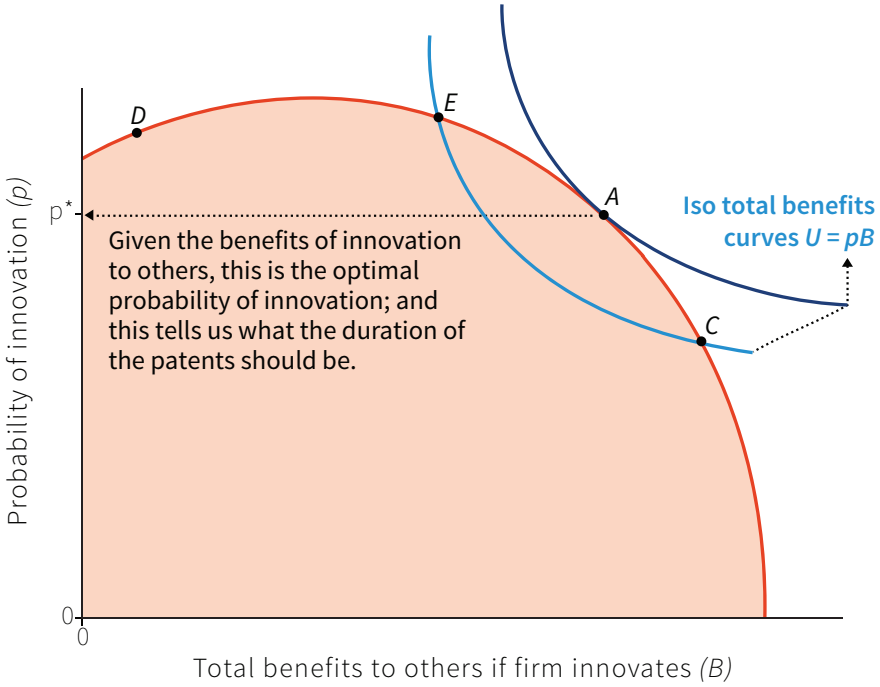
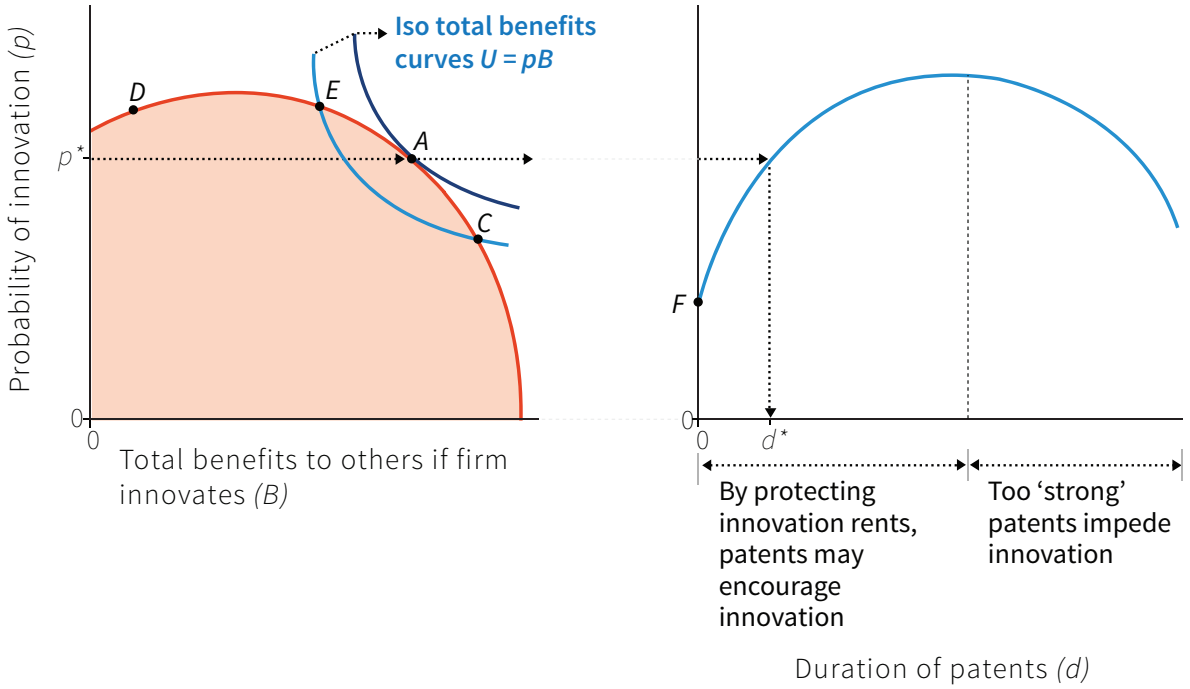


Figure 20.15 *The optimal probability of innovation for society.*

This outcome on its own is not a policy, but it allows us to determine one. We can now go back to Figure 20.13 and ask what patent duration would a policymaker set so that the innovating firm will choose society's optimal probability of innovation, p^*? Figure 20.16 shows the answer.



The optimal duration of patents

If we know p^*, we can use Figure 20.13 (the right-hand figure here) to determine the optimal duration of patents, d^*.

What if there were no patents?

At point *F* there are no patents. We can see that innovation will still occur, but below the optimal level for society.

Figure 20.16 *The optimal patent duration.*

DISCUSS 20.7: OPTIMAL PATENTS

1. Choose two contrasting technologies. For one, the government would optimally choose a short patent duration. For the other, it would choose a longer patent duration. In each case, draw the feasible set.
2. The length of patents and copyrights has increased steadily since the Industrial Revolution. Does this necessarily reflect the move to an optimal patent duration over this period?
3. How should patent offices react if firms seek to cement patent monopolies by patenting improved versions of the original technology at a later date? (A process known as *evergreening*.)

20.8 PUBLIC FUNDING OF BASIC RESEARCH, EDUCATION, AND INFORMATION INFRASTRUCTURE

The pros and cons of various kinds of intellectual property rights are just a part of the problem of designing an effective innovation system. Another important element is the role of the government. The origins of the computer and, by extension, the entire information revolution make this clear.

Government-funded research

The roots of the IT revolution can be traced to the building of the first electronic programmable computers after the second world war, although as with any technology, some elements are older. Charles Babbage first proposed a computer in the shape of his Difference Engine in a learned paper published in 1822 (and was funded by the British government to develop it).

The British and American governments' efforts during and after the second world war pioneered programmable electronic computing. In the US, the early focus was on the development of missile systems and then the Manhattan Project to develop the atomic bomb, and the need in both ballistics and predicting atomic reactions for huge numbers of rapid calculations. US government money supported private entities such as Bell Labs in New Jersey, as well as Federal research facilities like Los Alamos.

There was a close partnership between the private sector, government agencies and universities, resulting in the building of the ENIAC machine in 1946 under the auspices of the US Army. It was the first electronic computer, although it could not store programmes. Other innovations followed swiftly, such as the development of the transistor by William Shockley at Bell Labs in 1948, as well as the creation of new companies such as Fairchild Semiconductor. American government support for the industry has continued through research funding, including, famously, the creation of the internet (in 1969) in a project financed by the Defense Advance Research Projects Agency, or DARPA.

In the UK early progress in computing was focused on the efforts at Bletchley Park, where the mathematician Alan Turing worked, to crack Germany's Enigma code. The Colossus machine developed there remained a secret until the 1970s, but Bletchley Park scientists and engineers went on to build in 1948 the world's first post-war stored programme computer with a memory, called Baby, at the University of Manchester, another publicly funded institution. The commercial development of computers followed swiftly, by companies such as Ferranti.

This pattern of government funding of early-stage research, either through government agencies including the military or through universities, followed by commercial applications is common. As well as the computer and electronics industries, the internet, and the World Wide Web (created by Tim Berners-Lee at the CERN research laboratory funded by a consortium of governments), the modern pharmaceuticals and biotech sectors, and commercial applications of new materials such as graphene all have roots in publicly financed basic research and early-stage development. Touch screens and the computer mouse were also the result of US government-funded research.

The MP3 format was created by a small group of researchers at a public research lab in Germany, belonging to the Fraunhofer Gesellschaft. Their patent allows shrinking the size of audio files by a factor of 12, while maintaining sound quality. This innovation made music sharing via the internet possible and contributed to major upheaval in the global music industry. Commercial firms did not initially adopt it as a standard, and it became widely diffused because the creators responded by distributing encoding software to users for a low price and did not pursue hackers who then made it available for free.

In this video, Mariana Mazzucato, an economist who specialises in the causes and impacts of innovation, uses the example of some of the basic digital innovations such as the internet, GPS and touch screens to argue that the government has an essential role in funding research and start-up technology companies. She sees the government's role not just as filling in activities the market will not undertake, perhaps because the returns are too far in the future and uncertain, but also as shaping what kind of activities the private sector will do. Strategic investment by the US government helps explain why American companies dominate high-tech industries including digital and biotechnology.

DISCUSS 20.8: GOVERNMENT-FUNDED RESEARCH

In this video, Mazzucato suggests that governments should start to take investment stakes in technology companies, so they earn a return on the funds they invest in research.

1. What are the arguments for and against direct government investment in the commercial application of new technologies?
2. Describe ways in which governments could pick technologies in which to invest, so that the process would be more transparent to taxpayers.
3. Do you think it would be sensible to involve taxpayers in the choices about which technologies in which to invest? Explain your answer.
4. Which kind of technologies do you think governments should spend more on and which technologies should governments leave to the private sector? Explain your answer.

Competitions and prizes

A quite different policy for the support of innovation is to award a prize for the successful development of a solution to a problem that will meet some specifications. The prizewinner is rewarded for the cost of development, rather than with a monopoly over the novel idea or method, and the innovation then goes immediately into the public domain.

For example, in the aftermath of the Deepwater Horizon oil rig disaster (shown in the title photo of Unit 18) the XPrize Foundation offered $1m to any team who could significantly improve current technology for the cleanup of oil spills. Within a year a team had devised a method that quadrupled the industry-standard recovery rate.

A more famous example is the invention by watchmaker John Harrison of the marine chronometer, a device which for the first time allowed the (reasonably) accurate measurement of a vessel's longitude at sea. Harrison started work on his chronometer in 1730 in response to an offer made in 1714 by the British government of a cash prize (about £2.5m in 2014 prices) for the invention of a device to measure longitude. Harrison's approach to the challenge was to build an accurate clock small enough to be seaborne in order that the Greenwich time at which the sun reached its zenith could be determined. This would allow the ship's position west of Greenwich to be calculated. The problem had attracted some of the best minds of the time, including Isaac Newton's. Harrison produced many versions, each better than the last, but argued with the government about whether he deserved the prize money. The argument arose because Harrison's solution to the problem was rather different from that expected by the government. He was awarded a series of smaller sums over the years.

Another example of where the competitions work well is the creation of prizes for the successful development of drugs for neglected diseases. These drugs treat illnesses that are common in parts of the world in which there is little pharmaceutical innovation because the private market for them is small.

20.9 SLOWER PRODUCTIVITY GROWTH IN SERVICES, AND THE CHANGING NATURE OF WORK

Adequate public policies concerning innovation, as we have seen, can help in two main ways:

- *Increasing the pace of innovation:* This occurs through such interventions as the support of basic research and communications infrastructure, standard-setting, as well as the design of patents, copyright and trademarks.

- *Influencing the direction of innovation:* This tilts the process towards the production of novel ideas and applications with environmental, learning, medical or other socially valued applications.

Now we can add a third role for a public innovation policy:

- *Addressing the consequences of innovation:* This includes its interaction with other long-run trends for the workings of the entire economy and our quality of life.

Here and in the next section we examine how public policy can influence four of the consequences of innovation:

- *A shift from goods manufacturing to services:* There has been a greater increase in productivity in the sectors of the economy that produce goods than in services, and as a result employment has shifted from the production of goods to the production of services.
- *Services produced in the market:* The provision of many services has shifted from home production to work for pay.
- *The rise of the robots:* The contrasting fortunes of workers with machine-replaceable skills and endowments on the one hand, and those for who machines increase the value of their endowments on the other.
- *Productivity growth slows:* If service productivity grows more slowly than manufacturing productivity, the shift from goods to services reduces overall productivity growth in the economy.

The rise and fall of manufacturing employment

Before the Industrial Revolution (as you know from Unit 1), most of the output of the economy was made by family members. They were not employees but instead were independent producers of the goods and services not only for their own use, but also for sale to others. The Industrial Revolution and the emergence of the capitalist economic system shifted labour from the family to the firm: independent producers became employees.

Due to technological progress in machine-based production, manufactured goods became cheaper. As a result, textiles and clothing once produced in the home were now purchased and paid for with the wages gained through industrial and other employment. The result was a sustained

INDUSTRY

Data is collected by national statistical authorities on business activities and classified in two broad categories: services and goods.

- *Service industries* comprise businesses whose principal activity is to provide services rather than tangible products.
- *Goods-producing industries*, often called “industry”, are agriculture, mining, manufacturing and construction. Manufacturing is the most important component of “industry”.

increase in employment in the industrial sector of the economy. Manufacturing makes up most of employment in *industry* and the terms manufacturing and industry are often used interchangeably.

This permanent technological revolution had two effects:

- *The employee share of the gains increased:* As their bargaining power increased, employees (as you saw in Unit 2) were able get a larger share of the productivity gains made possible by the advance in technology.
- *Fewer farmers in the labour force:* Farming became more productive. As people became richer they spent less of their budget on food. Therefore the fraction of the labour force engaged in farming fell.

For many the shift out of farming and the rise of manufacturing employment meant an improvement in economic opportunities, especially when the trade unions and worker-based political parties forced employers to improve industrial working conditions.

This did not last for ever. Figure 20.17 shows that for most of the world's large economies, the era of expanding manufacturing employment ended sometime in the third quarter of the 20th century. Just as manufacturing had initially displaced agriculture as the main kind of employment, the production of services rather than goods has replaced manufacturing.

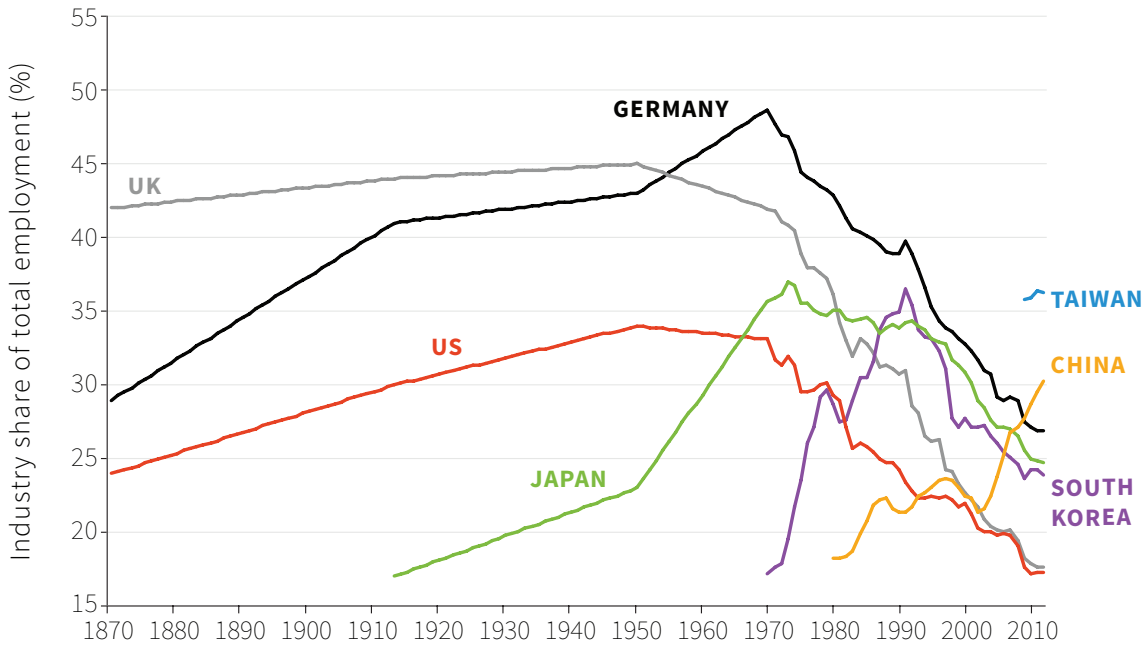


Figure 20.17 *The rise and fall of the share of employment in industry (1870-2012).*

Source: *US Bureau of Labor Statistics. 2012. "International Labor Comparisons (1970-2012)." International Labour Association. 2015. "ILOSTAT Database."*

You can see from the figure that the shift of employment out of industry was led by the early technological leaders, the UK and the US, followed by Japan and Germany. Even South Korea, whose meteoric rise to industrial prominence began only in the last quarter of the 20th century, saw its share of manufacturing employment begin to decline before the end of the 20th century.

There are two exceptions to this rise-and-fall story. China and Taiwan now have a larger share of their labour force in industry than Germany, which is the rich country where manufacturing employment has remained highest. In both China and Taiwan, labour continued to be pulled into the manufacturing sector in the 21st century, in part to satisfy the demand for their goods in world markets.

The economics of slower productivity growth in services

The amount of labour devoted to agriculture has declined in all of the countries shown in the figure. Fewer than one in 20 workers in rich countries work in agriculture. The big recent shift in work has been from the production of goods (manufacturing and agriculture) to the production of services. We know that output per hour of labour (productivity) is growing more slowly in the production of services than in manufacturing. This has two effects:

- To produce the same mix of goods and services it now takes relatively less labour devoted to goods and more to services.
- The costs of producing goods have fallen relative to the costs of producing services, and so the prices of goods have fallen relative to the prices of services, leading people to buy more goods and fewer services than they otherwise would have done.

The first of these effects has been stronger than the second.

To see how this process works, let's simplify using a model in which only the first effect occurs. So we assume that people consume a given ratio of goods (shirts, say) and services (haircuts). The examples illustrate the reason for slower productivity growth in services: it takes about as long today to cut someone's hair as it did 100 or even 200 years ago; to produce a shirt it takes much less time, probably less than a fifth of the time it did 200 years ago.

Figure 20.18 shows the model. The total amount of labour employed in the economy is assumed to be 1 (it could be 1 million hours, for example). If all of this labour is devoted to the production of goods, 1 unit of goods is produced. And the same is true of services: if all the labour produces services, then 1 unit of services is produced.

The solid red line is the feasible frontier, showing the amounts of goods and services that are possible given the existing technologies and the amount of labour employed. We assume that the same number of units of goods and services are consumed; in the figure the quantity of services and the quantity of goods consumed both equal 1/2 unit.

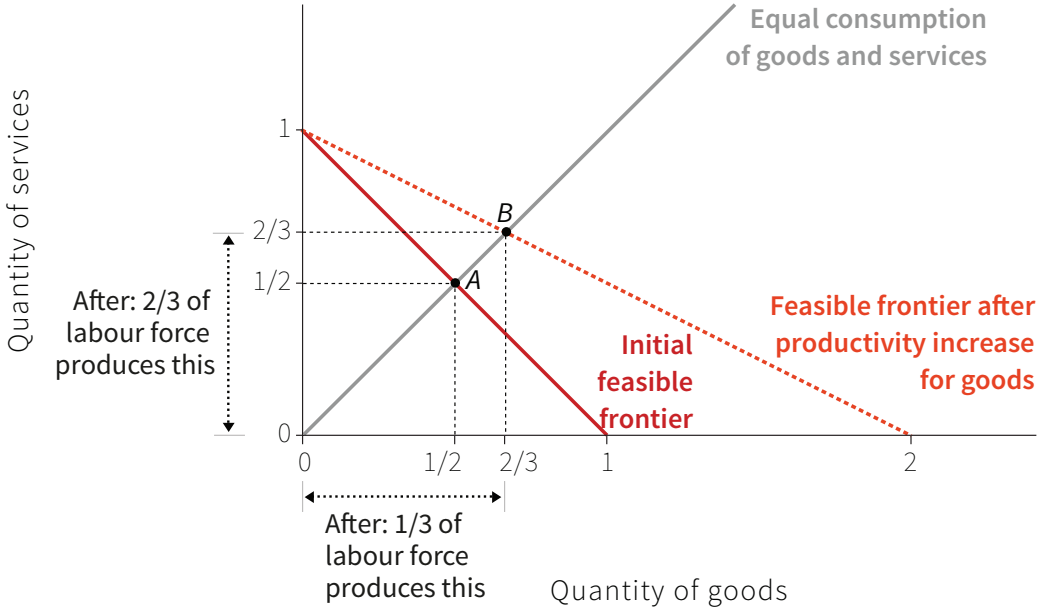


Figure 20.18 *Increased productivity in goods production raises the fraction of workers in services.*

Now consider what happens when the productivity of labour increases in the production of goods. (For simplicity we assume that there is no increase in productivity in services). If all of those employed now produce goods, the output will be 2 units rather than 1 unit. The new feasible frontier is shown as a dashed line in the figure. The fact that the feasible frontier is flatter than before illustrates the fact that goods are now less costly to produce (they require only half as much labour as before, and half as much as a unit of services).

If people continue to consume equal amounts of goods and services, the economy will be at point *B* on the new feasible frontier: with total output of 2/3 units of each. But now producing these equal amounts of goods and services does not require an equal amount of labour: only 1/3 of the labour is sufficient to produce 2/3 units of goods, while 2/3 of the labour is required to produce 2/3 units of services.

Labour has shifted from goods production to services production. The model is designed to illustrate why the shift took place. Two things left out of the model have, in reality, reduced the shift, and a third one has increased it:

- *Productivity increases in some services reduce the shift:* We assumed that there was no productivity increase in services. But think of the kinds of services we have discussed in this unit such as the sharing of music or other forms of digital information where the productivity advances have been large. If productivity in services increased, then in our model it would at least partially offset the shift in labour. We will see just below, however, that much of the service sector of the economy is made up of such things as personal care, which is more like haircuts than the reproduction of music.

- *Substitution of goods for services reduces the shift:* We increase the proportion of goods we consume if their relative price falls. By assuming that the ratio of goods (shirts) to services (haircuts) did not change, we ignored this process. It would partially offset the decline in goods employment.
- *An increase in relative demand for services increases the shift:* We also ignored the possibility that as incomes rise, people choose to spend more of their budget on services. Remember that services include tourism and other forms of recreation, and also include health, education and care, which may not be paid directly out of the household's disposable income. This would reinforce the shift of labour into services. We have seen this before: it is equivalent to the earlier shift of labour out of agriculture that occurred when the share of food in household budgets shrank.

But, in the countries showing a decline in goods employment relative to services, the net effect of the things we have excluded from the model did not completely offset the deindustrialisation of the workforce.

Some countries import many of the goods they consume, while others export much of the output of goods-producing workers. This is relevant too: it helps explain why different countries have different patterns for the hump-shaped relationship shown in Figure 20.17. International trade and the opportunities for specialisation that came with it accelerated the decline in the goods-producing share of employment in some countries (the US and the UK, for example), but retarded it in others (Germany, South Korea). China's growing share of employment in goods reflects both the forces seen elsewhere in the now-rich countries as well as its specialisation in exporting manufactures. The logic behind Figure 20.18 and the analysis of the result of a productivity increase in goods production is illustrated in this unit's Einstein section.

20.10 INNOVATION AND INEQUALITY

The declining share of employment dedicated to producing goods and the increasing share producing services has been accompanied in many high-income economies by two additional trends:

- *Services outside the home:* Services once produced almost exclusively in the family (food preparation, caring for the old, teaching the young, and curing the sick) have been increasingly performed by employees outside the home working in schools, restaurants, day care centres, elderly care establishments and hospitals.
- *Innovation has unequal effects:* Innovation affects people differently, depending on whether their skills and other endowments could be replicated by machines (*human replacement*, or *automation*), or whether machines make their endowments more valuable.

The missing middle in the United States

The data in Figure 20.19a illustrates both trends for the US economy. We have used the US economy as an illustration because of the quality of the available data, but similar trends are evident in other high-income countries.

Figure 20.19a arranges jobs from the highest paid (in hourly wages) at the top to the least well paid jobs at the bottom, and estimates growth or contraction of employment on the horizontal axis.

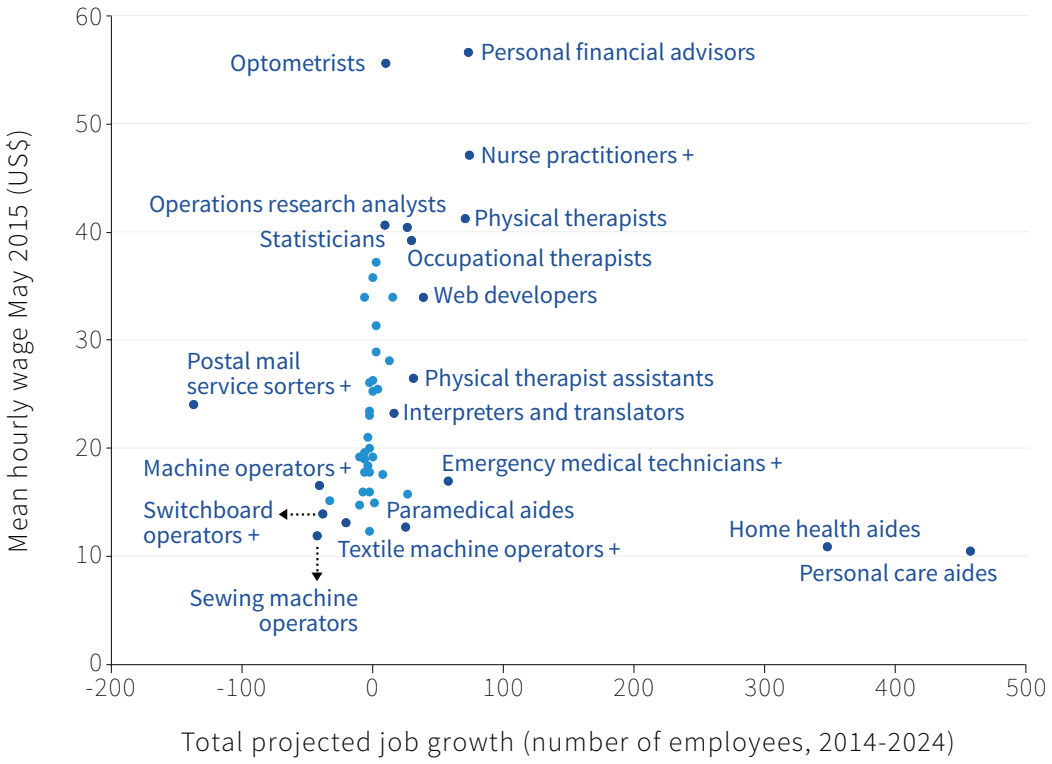


Figure 20.19a *The missing middle in the US (2014-2024): Occupations forecast to undergo 20% or more job change. Note: "+" indicates similar occupations.*

Source: *US Bureau of Labor Statistics. 2014. "Employment Projections." US Bureau of Labor Statistics. 2015. "Occupational Employment Statistics."*

Notice these things about the data:

- *Missing middle:* Both high-wage and low-wage occupations are adding many jobs, but employment gains among the occupations with wages in between are more limited.
- *Workers substitute for families:* The biggest increases are in human services, most of them in health-related professions; the biggest gainers by far are occupations that substitute for work once done primarily by family members.
- *Routine work is done by machines:* Digitalisation reduces the demand for routine tasks, while increasing demand for low- and high-skilled works.

The high-wage job gainers not in human services are all occupations (operations researchers, statisticians and web developers) in which digital information processing has greatly increased the productivity of workers with the right skill endowments.

- *The average workers are the losers:* Occupations with job losses tend to have near average wages or less. In these cases new technologies allow machines to do routine work once performed by workers such as postal carriers, industrial machine operators and switchboard operators.

Figure 20.19a showed only occupations for which gains or losses are projected to be at least 20% of 2014 employment. But as Figure 20.19b shows, this pattern holds when we look at all jobs in the US economy.

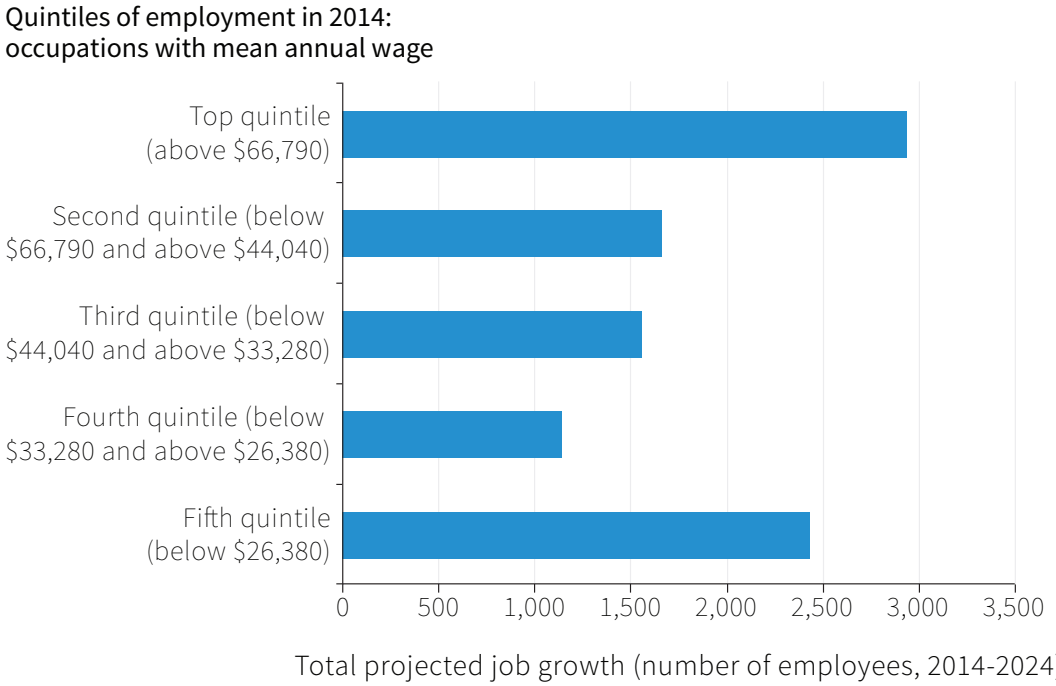


Figure 20.19b *The missing middle in the US (2014-2024): Job growth is highest in the top fifth and bottom fifth of occupations in the US, by annual earnings.*

Source: *US Bureau of Labor Statistics. 2014. "Employment Projections." US Bureau of Labor Statistics. 2015. "Occupational Employment Statistics."*

The projected trends shown in figures 20.19a and 20.19b have been underway in the US since at least the 1970s.

Labour-replacing and labour-enhancing innovation: Effects on inequality

It is clear from Figures 20.19a and 20.19b that these changes in the distribution of jobs will affect the degree of income inequality in the economy as a whole. We can study these effects using the Lorenz curve and the Gini coefficient derived from it.

To see how, consider a hypothetical economy in Figure 20.20, before and after it introduces machines to perform routine operations that had always been done by humans. We'll call these machines robots. The solid blue Lorenz curve depicts the distribution of income between five employers and 95 workers before the introduction of the robots. Five of the workers are unemployed; and among the 90 who are employed, all receive the same wage, whether they do routine or non-routine work.

The slope of the flatter of the two upward-sloping lines is an indication of how much workers are paid: we see that the 90 employed workers receive 60% of the income of the economy. So each receives 60/90%, or 2/3% of what the economy produces. The slope of the steeper solid line shows that five owners receive 40% of the income, so that each receives 8% of the output of the entire economy.

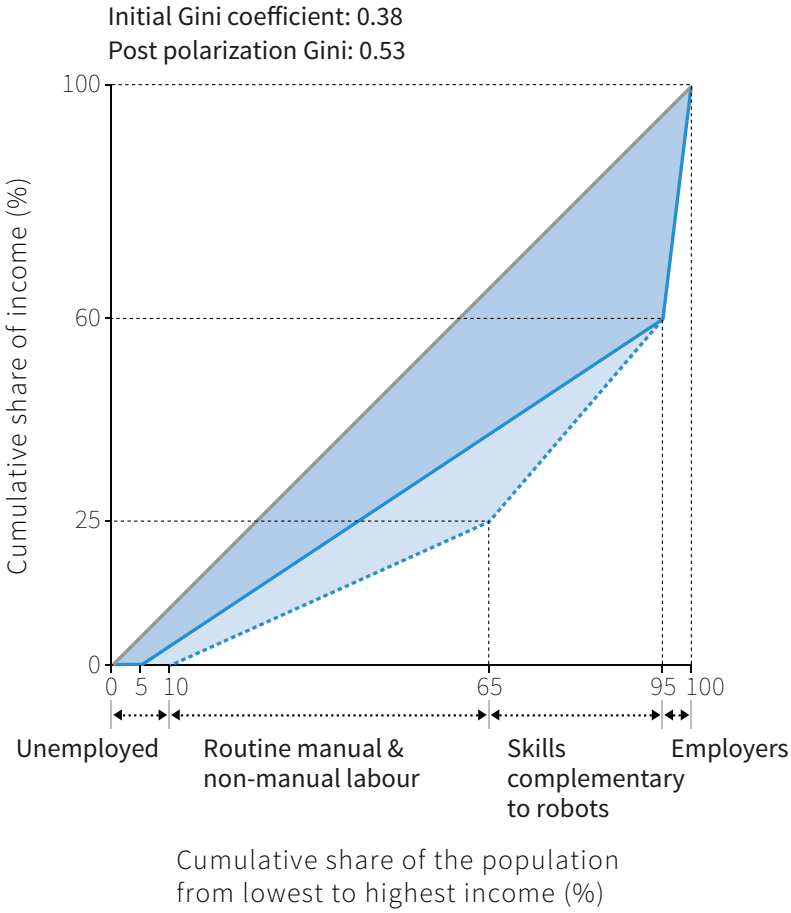


Figure 20.20 *The effect of robots on inequality.*

To understand the impact in the short run of the plan to introduce robots, think about the skill endowments of the workers. Sixty of them are doing routine jobs that were once relatively well-paid—such as machine tending or mail sorting—and that can now be done by robots. Others have the training to not just operate machinery, but to design, repair and calibrate machinery and manage its deployment.

We now study the impact of the introduction of robots:

Short-run effects

- *The robots are labour-replacing:* For routine jobs in which the machines and skills are *substitutes*, the value of a worker's endowment is reduced by the new technology because the robot can replace the worker.
- *The robots are labour-enhancing:* For those jobs in which the machines and skills are *complements*, the value of a worker's endowment is increased by the new technology.

These two effects are shown in the new (dashed) Lorenz curve depicting the short-run effects of the new technology on workers who previously earned two-thirds of a percent of output each. At least some of the 60 workers for whom the robots are labour-replacing lose their job. Five of them have now joined the unemployed; their labour has been replaced by the machines. Those who remain employed have suffered a fall in their bargaining power (because they too can be replaced). These 55 workers now receive 25% of the output of the economy, and their earnings fall to 0.5% of the total output each.

On the other hand, the 30 workers with skills that are complementary to the robots have gained. They now receive 35% of the output of the economy, or a little more than 1% each.

The result—shown by the new Lorenz curve falling farther below the perfect equality line—is that the Gini coefficient increases from 0.38 to 0.53.

Long-run effects

In the long run we may see an increase in profits: the five owners still receive 40% of the economy's output, but due to the increase in labour productivity the economy now produces more, so their profits have increased, providing them with an incentive to invest more.

Now we use an analysis similar to the one we did in Unit 15:

- *The labour-replacing process of innovation may displace labour:* This sends workers into unemployment.
- *This process generates profits:* In the long run they motivate and finance an expansion of the capital stock of the economy.
- *This creates additional employment.*

Taxation levied on the enhanced profits of the owners and the workers with increased wages can be used to finance:

- *Additional employment and opportunities for career progression and rising wages:* These opportunities would be in human services such as health and care, where jobs are non-routine but often poorly paid.
- *Opportunities for workers with routine skills to upgrade their endowments:* Their labour becomes machine-enhanced rather than machine-replaceable; for example a former drill press operator learning how to code.

We know that the long run need not be one of joblessness and greater inequality. We have been worried that machines will supersede working people since at least 1811, but there are clearly many examples of technology leading to increased profitability and higher demand, even as some workers become unemployed.

One example is the introduction of Automatic Teller Machines (ATMs) by banks. Surely this would have increased unemployment among human bank tellers? James Bessen, an economist, looked at the employment levels in the US and found that the number of bank tellers continued to rise even after the machines were installed. Rather than mechanical tasks, they were now used to provide other services such as advice to customers. Similarly, Bessen found employment increased among book-keepers and retail sales staff despite automation of some of their tasks; but on the other hand technology did displace jobs for travel agents. The complementarity of workers' skills and the expansion of demand as a result of innovation determine what happens to employment.

WHEN ECONOMISTS DISAGREE

THE END OF THE PERMANENT TECHNOLOGICAL REVOLUTION?

We began Unit 1 with the Industrial Revolution, the capitalist revolution, and history's hockey sticks of rapid technological progress. In Unit 2 we explained how these advances translated into improvements in wellbeing. And we began this Unit with the dramatic (and possibly even accelerating) rate of technical advance in computation.

The economy will be producing relatively more services, rather than goods, in the future. Will this limit the ability of technological progress to increase labour productivity at the rate that has occurred since the Industrial Revolution, and especially during the golden age of capitalism? It seems appropriate to end a unit about Innovation with disagreement among economists about whether the "permanent" technological revolution is ending.

Figure 20.21 *The growth rate of productivity over the long run.*

Source: *Bolt, Jutta, and Jan Juiten van Zanden. 2013. "The First Update of the Maddison Project Re-Estimating Growth Before 1820." Maddison-Project Working Paper WP-4, January. Broadberry, Stephen. 2013. "Accounting for the Great Divergence." London School of Economics and Political Science. The Conference Board. 2015. "Total Economy Database."*

Figure 20.21 shows the best available data on the advance of productivity of labour in the UK since 1400, and also for the US for the period in which the US has been global technology leader. Robert Gordon, an economist who specialises in productivity and growth, has written extensively about productivity growth and its effects. Follow this link to read the first chapter of his book *The Rise and Fall of American Growth*.

Gordon argues that the rapid growth era from in the first half of the 20th century is long gone, and slower growth lies ahead of us. In contrast, Erik Brynjolfsson and Andrew McAfee, both economists, argue that digital technology is opening up a "second machine age". Watch these two videos (video 1 and video 2), in which they explain their point of view.

DISCUSS 20.9: THE PERMANENT TECHNOLOGICAL REVOLUTION

Use all the sources above, plus these two newspaper articles (article 1 and article 2) to answer the following questions:

1. According to Gordon, Brynjolfsson and McAfee, which other factors, apart from technological innovation, affect the trend in the rate of GDP per capita growth? Why might it take a long time for today's innovations to affect the economy's growth rate?
2. Would you agree that modern inventions such as mobile internet access are less important than 20th century inventions such as indoor plumbing? Explain why or why not. In India, many people have access to mobile phones but not indoor toilets. Is this consistent with your argument? Explain.
3. Assess how well GDP per capita growth measures the effect of innovation. Discuss how hedonic pricing and the contingent valuation methods introduced in Unit 18 could be used to estimate the value of the improved quality of goods resulting from an innovation.
4. Use the evidence and models from Units 2, 15 and 20 to discuss whether you agree with the analysis by Brynjolfsson and McAfee of the relationship between technological progress and inequality in each machine age.

20.11 CONCLUSION

The UK and the Netherlands, birthplaces of capitalism and the Industrial Revolution, were not unique in the intelligence and creativity of their peoples. China, arguably, had proven to be an equally if not more inventive society in earlier years having first developed paper, printing, gunpowder, the compass, and literally hundreds of other important innovations. Other countries, notably Japan, were adept at the adaptation and spread of novel methods and ideas. But the combined pull of innovation rents and the push of competition to survive characteristic of the innovation and diffusion process under capitalism made it a uniquely dynamic economic system.

Public policy also played an important part. For innovators to take the risk of introducing a new product or production process, it is crucial for their innovation rents not to be seized by the government or by others. This requires that property rights are protected by a well-functioning legal system. Silicon Valley, the German innovation system, or other successful examples of innovation have often been assisted by governments that provide complementary inputs such as physical infrastructure and public education, and allow the innovator only a temporary monopoly so that competition eventually will reduce prices.

In a nutshell, it is this combination of private incentives and supportive public policy that explains why capitalism is such a dynamic economic system.

CONCEPTS INTRODUCED IN UNIT 20

Before you move on, review these definitions:

- *Process innovation* and *product innovation*
- *Radical innovation* and *incremental innovation*
- Innovations as *substitutes* or *complements*
- *Labour-replacing innovation* and *labour-enhancing innovation*
- *General-purpose technologies*
- *Innovation systems* (Silicon Valley and Germany)
- *First copy costs*
- *Winner-take-all competition*
- *Patents, Copyrights, Trademarks*
- *Demand-side economies of scale* and *network external effects*
- *Matching (two-sided) markets*
- *Optimal patent duration*

Key points in Unit 20

Innovation

Innovation has substantial external effects; it is characterised by the problem of coordination among innovators, the public goods nature of new knowledge, and important economies of scale leading to winner-take-all competition.

Successful innovation systems

Successful systems combine knowledge creation and sharing, private and public methods of coordinating innovative activity to internalise the external effects, and mix public and private funding sources.

Two-sided markets

These markets match individuals from two distinct sets of people. When they don't work well they illustrate the lock-in effects and market failures that are associated with network economies of scale and winner-take-all competition.

Intellectual property rights (IPR)

Patents, copyrights, trademarks and other intellectual property rights create temporary monopolies to allow economic profits (innovation rents) for successful innovators, but they are a barrier to diffusion and may retard invention.

Persistence of innovation rents

Innovation rents may also be protected by secrecy, or the time taken by competitors to copy an innovation.

Public policies

Non-IPR policies that support innovation include government-funded basic research; education to foster curiosity, capability, and creativity; standard setting and prizes.

The transformation of paid work

Rapid technical progress in goods production (by comparison to most services), and an increasing role for paid employment in the production of services, are transforming the nature of paid work.

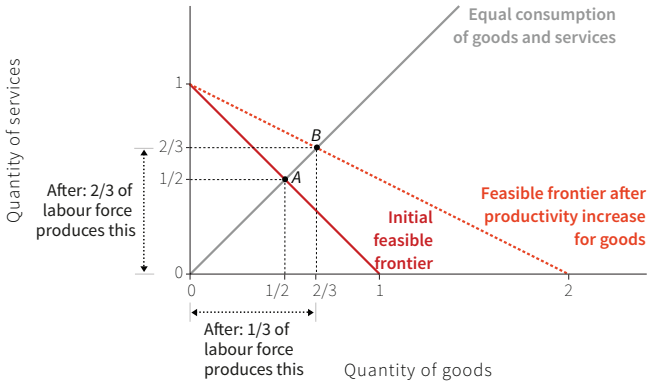
Innovation and inequality

New machines are substitutes for some workers' skill endowments, and are complements to others' endowments. This has contributed to inequality in earnings.

20.12 EINSTEIN

How faster productivity growth in goods production may shift employment from goods to services

This Einstein explains the logic behind Figure 20.18 (right) and explains why a productivity increase in goods production shifts employment to firms that produce services. We define q_s as the productivity of labour in services. Then $q_s = Q_s/L_s$, the quantity of services divided by the amount of labour employed to produce it. Look at these equations:



$$Q_s L_s = Q_s = Q_g = q_g L_g$$

- $Q_s L_s = Q_s$: The productivity of labour in services multiplied by the amount of labour in services is equal to the amount of services produced.
- $Q_s = Q_g$: The output of goods must be the same as the output of services. This isn't always true, but we defined it that way in our model.
- $Q_g = q_g L_g$: The output of goods is equal to the productivity of labour in the production of goods multiplied by the amount of labour employed in producing goods.

We can now equate the first and last terms of the above equation to give us an expression for the amount of labour that must be employed in the two sectors, given the productivity levels in each sector, if they produce an equal number of units of output:

$$Q_s L_s = q_g L_g$$

We then rewrite this using the fact that the total amount of labour in the two sectors sums to one:

$$Q_s L_s = q_g L_g = q_g(1 - L_s)$$

Then we rearrange the equation to get an expression for the amount of labour engaged in service production:

$$L_s = \frac{q_g}{q_g + q_s}$$

In the figure, productivity in both of the two sectors was 1, so the amount of labour engaged in the goods-producing service was 1/2. When the productivity of labour in goods production doubles:

$$L_s = \frac{2}{1+2} = \frac{2}{3}$$

This is the share of labour devoted to the production of services after the increase in productivity of the labour used in the production of goods.

20.13 READ MORE

Bibliography

1. Benkler, Yochai. 2006. *The Wealth of Networks: How Social Production Transforms Markets and Freedom*. New Haven, CT: Yale University Press.
2. Bessen, James. 2015. *Learning by Doing: The Real Connection between Innovation, Wages, and Wealth*. New Haven, CT: Yale University Press.
3. Boldrin, Michele, and David K Levine. 2008. *Against Intellectual Monopoly*. New York, NY: Cambridge University Press.
4. Boldrin, Michele, and David K Levine. 2013. "The Case against Patents." *Journal of Economic Perspectives* 27 (1): 3–22.
5. Bolt, Jutta, and Jan Juiten van Zanden. 2013. "The First Update of the Maddison Project Re-Estimating Growth Before 1820." *Maddison-Project Working Paper WP-4*, January.
6. Boseley, Sarah. 2016. "Big Pharma's Worst Nightmare." *The Guardian*, February 5.
7. Bowe, Christopher. 2011. "Say Farewell to Lipitor but Don't Forget Its Lessons." *Harvard Business Review*. November 18.
8. Broadberry, Stephen. 2013. "Accounting for the Great Divergence." London School of Economics and Political Science. November 1.
9. Cohen, Wesley M, Richard R Nelson, and John P Walsh. 2000. "Protecting Their Intellectual Assets: Appropriability Conditions and Why U.S. Manufacturing Firms Patent (or Not)." *NBER Working Paper 7552*. National Bureau of Economic Research.
10. Coyle, Diane. 2015. "Thinking, Learning and Doing." *Enlightenment Economics Blog*. October 23.
11. DiMasi, Joseph A, Ronald W Hansen, and Henry G Grabowski. 2003. "The Price of Innovation: New Estimates of Drug Development Costs." *Journal of Health Economics* 22 (2): 151–85.
12. Edsall, Thomas B. 2016. "Boom or Gloom?" *New York Times*, January 27.

13. Engel, Jerome S. 2015. "Global Clusters of Innovation: Lessons from Silicon Valley." *California Management Review* 57 (2). University of California Press: 36–65.
14. Fisher, William W. 2004. *Promises to Keep: Technology, Law, and the Future of Entertainment*. Palo Alto, CA: Stanford Law and Politics.
15. Gilbert, Richard. 2011. "A World without Intellectual Property? A Review of Michele Boldrin and David Levine's against Intellectual Monopoly." *Journal of Economic Literature* 49 (2): 421–32.
16. Giorcelli, Michela, and Petra Moser. 2015. "Copyright and Creativity: Evidence from Italian Operas." *SSRN Electronic Journal*. Social Science Electronic Publishing.
17. Gordon, Robert J. 2016. *The Rise and Fall of American Growth: The U.S. Standard of Living since the Civil War*. Princeton, NJ: Princeton University Press.
18. Graham, Stuart J. H, Robert P Merges, Pamela Samuelson, and Ted M Sichelman. 2009. "High Technology Entrepreneurs and the Patent System: Results of the 2008 Berkeley Patent Survey." *Berkeley Technology Law Journal* 24 (4): 255–327.
19. Hall, Bronwyn H, and Dietmar Harhoff. 2012. "Recent Research on the Economics of Patents." *Annual Review of Economics* 4 (1): 541–65.
20. Hall, Peter A, and David Soskice. 2001. *Varieties of Capitalism: The Institutional Foundations of Comparative Advantage*. New York, NY: Oxford University Press.
21. Hemphill, C. Scott, and Bhaven N Sampat. 2012. "Evergreening, Patent Challenges, and Effective Market Life in Pharmaceuticals." *Journal of Health Economics* 31 (2): 327–39.
22. International Labour Association. 2015. "ILOSTAT Database."
23. Janeway, William H. 2012. *Doing Capitalism in the Innovation Economy: Markets, Speculation and the State*. Cambridge: Cambridge University Press.
24. Jensen, Robert. 2007. "The Digital Provide: Information (Technology), Market Performance, and Welfare in the South Indian Fisheries Sector." *The Quarterly Journal of Economics* 122 (3): 879–924.
25. Kapczynski, Amy, and Aaron Kesselheim. 2016. "Government Patent Use: A Legal Approach to Reducing Drug Spending." *Health Affairs* 35 (5): 791–97.
26. Kornai, János. 2013. *Dynamism, Rivalry, and the Surplus Economy: Two Essays on the Nature of Capitalism*. Oxford: Oxford University Press.
27. Koromvokis, Lee. 2016. "Are the Best Days of the U.S. Economy Over?" PBS NewsHour. January 28.
28. Kovacic, William E, and Carl Shapiro. 2000. "Antitrust Policy: A Century of Economic and Legal Thinking." *Journal of Economic Perspectives* 14 (1): 43–60.
29. Kremer, Michael, and Rachel Glennerster. 2004. *Strong Medicine: Creating Incentives for Pharmaceutical Research on Neglected Diseases*. Princeton, NJ: Princeton University Press.
30. Landes, David S. 2000. *Revolution in Time*. Cambridge, MA: Harvard University Press.

31. Levin, Richard C, Alvin K Klevorick, Richard R Nelson, and Sidney G Winter. 1987. "Appropriating the Returns from Industrial Research and Development." *Brookings Papers on Economic Activity* 1987 (3): 783–831.
32. Moser, Petra. 2005. "How Do Patent Laws Influence Innovation? Evidence from Nineteenth-Century World's Fairs." *American Economic Review* 95 (4): 1214–36.
33. Moser, Petra. 2013. "Patents and Innovation: Evidence from Economic History." *Journal of Economic Perspectives* 27 (1): 23–44.
34. Moser, Petra. 2015. "Intellectual Property Rights and Artistic Creativity."*Voxeu. org*. November 4.
35. Moser, Petra, and Alessandra Voena. 2012. "Compulsory Licensing: Evidence from the Trading with the Enemy Act." *American Economic Review* 102 (1): 396–427.
36. Mowery, David C, and Timothy Simcoe. 2002. "Is the Internet a US Invention?—an Economic and Technological History of Computer Networking." *Research Policy* 31 (8-9): 1369–87.
37. Nordhaus, William D. 2007. "Two Centuries of Productivity Growth in Computing." *The Journal of Economic History* 67 (01).
38. OECD. 2016. "Automation and Independent Work in a Digital Economy." *Policy Brief on the Future of Work*.
39. Rizvi, Zain, Amy Kapczynski, and Aaron Kesselheim. 2016. "A Simple Way for the Government to Curb Inflated Drug Prices." *Washington Post*, May 12.
40. Roth, Alvin. 1996. "Matching (Two-Sided Matching)." Stanford University.
41. Rysman, Marc. 2009. "The Economics of Two-Sided Markets." *Journal of Economic Perspectives* 23 (3): 125–43.
42. Saxenian, AnnaLee. 1996. *Regional Advantage: Culture and Competition in Silicon Valley and Route 128*. Cambridge, MA: Harvard University Press.
43. Simcoe, Timothy. 2012. "Standard Setting Committees: Consensus Governance for Shared Technology Platforms." *American Economic Review* 102 (1): 305–36.
44. Swarns, Rachel L. 2001. "Drug Makers Drop South Africa Suit over AIDS Medicine." *New York Times*, April 20.
45. The Conference Board. 2015. "Total Economy Database."
46. *The Economist*. 2007. "To Do with the Price of Fish." May 10.
47. United Nations University. 2016. "Greater Access to Cell Phones than Toilets in India: UN - United Nations University." UNU Press Releases. April 14.
48. US Bureau of Labor Statistics. 2012. "International Labor Comparisons (1970-2012)."
49. US Bureau of Labor Statistics. 2014. "Employment Projections."
50. US Bureau of Labor Statistics. 2015. "Occupational Employment Statistics."
51. Witt, Stephen. 2015. *How Music Got Free: The End of an Industry, the Turn of the Century, and the Patient Zero of Piracy*. New York, NY: Viking.