lyryx with Open Texts

Principles of Economics

D. Curtis & I. Irvine

Lyryx Version
2021-A

Principles of Economics

an Open Text by Douglas Curtis and Ian Irvine

Version 2021 — Revision A

Principles of Economics is a combination of *Principles of Microeconomics* and *Principles of Macroeconomics*. Those texts are both written by Douglas Curtis and Ian Irvine and released with a CC BY-NC-SA license. This text begins with the original 15 chapters from *Principles of Microeconomics* followed by chapters 4 through 13 of *Principles of Macroeconomics*. Both original texts are available to download for free at `lyryx.com`.

Be a champion of OER!

Contribute suggestions for improvements, new content, or errata:

- A new topic
- A new example
- An interesting new question
- Any other suggestions to improve the material

Contact Lyryx at info@lyryx.com with your ideas.

Principles of Economics

an Open Text

Version 2021 — Revision A

Attribution

To redistribute all of this book in its original form, please follow the guide below:

The front matter of the text should include a “License” page that includes the following statement.

> This text is *Principles of Economics* by I. Irvine and D. Curtis and Lyryx Learning Inc. View the text for free at `https://lyryx.com/subjects/economics/principles-of-economics/`

To redistribute part of this book in its original form, please follow the guide below. Clearly indicate which content has been redistributed

The front matter of the text should include a “License” page that includes the following statement.

> This text includes the following content from *Principles of Economics* by I. Irvine and D. Curtis and Lyryx Learning Inc. View the entire text for free at `https://lyryx.com/subjects/economics/principles-of-economics/`.
> <List of content from the original text>

The following must also be included at the beginning of the applicable content. Please clearly indicate which content has been redistributed from the Lyryx text.

> This chapter is redistributed from the original *Principles of Economics* by I. Irvine and D. Curtis and Lyryx Learning Inc. View the original text for free at `https://lyryx.com/subjects/economics/principles-of-economics/`.

To adapt and redistribute all or part of this book in its original form, please follow the guide below. Clearly indicate which content has been adapted/redistributed and summarize the changes made.

The front matter of the text should include a “License” page that includes the following statement.

> This text contains content adapted from the original *Principles of Economics* by I. Irvine and D. Curtis and Lyryx Learning Inc. View the original text for free at `https://lyryx.com/subjects/economics/principles-of-economics/`.
> <List of content and summary of changes>

The following must also be included at the beginning of the applicable content. Please clearly indicate which content has been adapted from the Lyryx text.

> This chapter was adapted from the original *Principles of Economics* by I. Irvine and D. Curtis and Lyryx Learning Inc. View the original text for free at `https://lyryx.com/subjects/economics/principles-of-economics/`.

Citation

Use the information below to create a citation:

Author: I. Irvine and D. Curtis

Publisher: Lyryx Learning Inc.

Book title: Principles of Economics

Book version: 2021A

Publication date: March 15, 2021

Location: Calgary, Alberta, Canada

Book URL: `https://lyryx.com/subjects/economics/principles-of-economics/`

For questions or comments please contact editorial@lyryx.com

Principles of Economics

an Open Text by Douglas Curtis and Ian Irvine

Base Text Revision History

Current Revision: Version 2021 — Revision A

Edits and revisions have been completed by the authors together with the editorial staff at Lyryx Learning.
All new content (text and images) is released under the same license as noted above.

2021 A	• *Principles of Microeconomics* and *Principles of Macroeconomics*, 2021A versions, have been combined to create this *Principles of Economics* combined textbook.

Table of Contents

About the Authors

Doug Curtis is a specialist in macroeconomics. He is the author of numerous research papers on fiscal policy, monetary policy, and economic growth and structural change. He has also prepared research reports for Canadian industry and government agencies and authored numerous working papers. He completed his PhD at McGill University, and has held visiting appointments at the University of Cambridge and the University of York in the United Kingdom. His current research interests are monetary and fiscal policy rules, and the relationship between economic growth and structural change. He is Professor Emeritus of Economics at Trent University in Peterborough, Ontario, and also held an appointment as Sessional Adjunct Professor in the Department of Economics at Queen's University in Kingston, Ontario from 2003 until 2013.

Ian Irvine is a specialist in microeconomics, public economics, economic inequality and health economics. He is the author of numerous research papers in these fields. He completed his PhD at the University of Western Ontario, has been a visitor at the London School of Economics, the University of Sydney, the University of Colorado, University College Dublin and the Economic and Social Research Institute. His current research interests are in tobacco use and taxation, and Canada's Employment Insurance and Welfare systems. He has done numerous studies for the Government of Canada, and is currently a Professor of Economics at Concordia University in Montreal.

Our Philosophy

Principles of Economics focuses upon the material that students need to cover in a first introductory course. It is slightly more compact than the majority of principles books in the Canadian marketplace. Decades of teaching experience and textbook writing has led the authors to avoid the encyclopedic approach that characterizes the recent trends in textbooks.

Consistent with this approach, there are no appendices or 'afterthought' chapters. No material is relegated elsewhere for a limited audience; the text makes choices on what issues and topics are important in an introductory course. This philosophy has resulted in a combined book of 15 Microeconomic chapters and 13 Macroeconomic chapters.

Examples are domestic and international in their subject matter and are of the modern era – consumers buy iPods, snowboards and jazz, not so much coffee and hamburgers. Globalization is a recurring theme.

While this book avoids calculus, and uses equations sparingly, it still aims to be rigorous. In contrast to many books on the market, that simply insert diagrams and discuss concepts in a diagrammatic framework, our books almost invariably analyze the key issues in each chapter by introducing a numerical example or case study at the outset. Students are introduced immediately to the practice of taking a set of data, examining it numerically, plotting it, and again analyzing the material in that form. This process is not difficult, but it is rigorous, and stresses that economics is about data analysis as well as ideas and theories. The end-of-chapter problems also involve a considerable amount of numerical and graphical analysis. A small number of problems in each chapter involve solving simple linear equations (intersecting straight lines); but we provide a sufficient number of questions for the student to test his or her understanding of the material without working through that subset of questions.

Structure of the Text

Principles of Economics combines the textbooks *Principles of Microeconomics* and *Principles of Macroeconomics*.

Principles of Microeconomics provides a concise, yet complete, coverage of introductory microeconomic theory, application and policy in a Canadian and global environment. Our beginning is orthodox: We explain and develop the standard tools of analysis in the discipline.

Economic policy is about the well-being of the economy's participants, and economic theory should inform economic policy. So we investigate the meaning of 'well-being' in the context of an efficient use of the economy's resources early in the text.

We next develop an understanding of individual optimizing behaviour. This behaviour in turn is used to link household decisions on savings with firms' decisions on production, expansion and investment. A natural progression is to explain production and cost structures.

From the individual level of household and firm decision making, the text then explores behaviour in a variety of different market structures from perfect competition to monopoly.

Markets for the inputs in the productive process – capital and labour – are a natural component of firm-level decisions. But education and human capital are omnipresent concepts and concerns in the modern economy, so we devote a complete chapter to them.

The book then examines the role of a major and important non-market player in the economy – the government, and progresses to develop the key elements in the modern theory of international trade.

Principles of Macroeconomics provides complete, concise coverage of introductory macroeconomic theory and policy. It examines the Canadian economy as an economic system, and embeds current Canadian institutions and approaches to monetary policy and fiscal policy within that system. Particular attention is given to the recent structure, performance, and evolution of the Canadian economy, and to the current targets and instruments of Canadian monetary and fiscal policy.

These are exciting and challenging times in which to study macroeconomics. We focus on short-run macroeconomic performance, analysis, and policy motivated by the recessions of the early 1980s and 1990s, the financial crisis and recession of 2008-2009, and the prolonged recovery that is still incomplete in several industrial countries in 2017. The early effects of the novel coronavirus, the COVID-19 pandemic, on employment, income, output and prices provided a whole new set of challenges for economic analysis and policy. To that end, the text examines macroeconomic institutions, performance, and policies in ways that help students understand and evaluate critically the news media coverage and broader public discussion of:

- Recessions and recoveries, unemployment, inflation, deflation and conditions in financial markets—topics of ongoing reporting, discussion, and debate.
- Monetary and fiscal policy announcements and discussions focused on inflation targets, interest rate settings, budget balances, tax rates, expenditures, and public debt targets as these affect economic performance.
- Exports, imports, international capital flows, foreign exchange rates, commodity prices and the importance of the international sector of the Canadian economy.
- Economic growth, productivity growth, and the importance of productivity growth for standards of living in Canada and other countries.

A basic modern Aggregate Demand and Supply model of real GDP and the *inflation rate* is developed based on:

- Expenditure decisions by households and businesses in an open economy.
- Government sector expenditures, taxes and budgets.
  - Current Canadian monetary policy based on inflation targets, interest rate policy instruments, and current Bank of Canada operating techniques, including the potential for quantitative or credit easing.
  - Canadian fiscal policy based on deficit and debt ratio targets, the government's budget function, the temporary shift to fiscal stimulus in 2009, the subsequent fiscal austerity designed to achieve a balanced budget by 2015 and the implications for economic performance and the public debt.

Numerical examples, diagrams, and basic arithmetic are used in combination to illustrate and explain economic relationships. Students learn about: The importance of consumption, capital expenditures, and government budgets; money supply; financial asset prices, yields, and interest rates; employment and unemployment; and other key relationships in the economy. Canadian and selected international data are used to provide real world examples and comparisons.

Part One

The Building Blocks

1. Introduction to key ideas
2. Theories, models and data
3. The classical marketplace – demand and supply

Economics is a social science; it analyzes human interactions in a scientific manner. We begin by defining the central aspects of this social science – trading, the marketplace, opportunity cost and resources. We explore how producers and consumers interact in society. Trade is central to improving the living standards of individuals. This material forms the subject matter of Chapter 1.

Methods of analysis are central to any science. Consequently we explore how data can be displayed and analyzed in order to better understand the economy around us in Chapter 2. Understanding the world is facilitated by the development of theories and models and then testing such theories with the use of data-driven models.

Trade is critical to individual well-being, whether domestically or internationally. To understand this trading process we analyze the behaviour of suppliers and buyers in the marketplace. Markets are formed by suppliers and demanders coming together for the purpose of trading. Thus, demand and supply are examined in Chapter 3 in tabular, graphical and mathematical form.

Chapter 1

Introduction to key ideas

In this chapter we will explore:

1.1 What's it all about?

The big issues

Economics is the study of human behaviour. Since it uses scientific methods it is called a social science. We study human behaviour to better understand and improve our world. During his acceptance speech, a recent Nobel Laureate in Economics suggested:

> *Economics, at its best, is a set of ideas and methods for the improvement of society. It is not, as so often seems the case today, a set of ideological rules for asserting why we cannot face the challenges of stagnation, job loss and widening inequality.*
>
> Christopher Sims, Nobel Laureate in Economics 2011

This is an elegant definition of economics and serves as a timely caution about the perils of ideology. Economics evolves continuously as current observations and experience provide new evidence about economic behaviour and relationships. Inference and policy recommendations based on earlier theories, observations and institutional structures require constant analysis and updating if they are to furnish valuable responses to changing conditions and problems.

Much of today's developed world still faces severe challenges as a result of the financial crisis that began in 2008. Unemployment rates among young people are at historically high levels in several economies, many government balance sheets are in disarray, and inequality is on the rise. In addition to the challenges posed by this severe economic cycle, the world simultaneously faces structural upheaval: Overpopulation, climate change, political instability and globalization challenge us to understand and modify our behaviour.

These challenges do not imply that our world is deteriorating. Literacy rates have been rising dramatically in the developing world for decades; child mortality has plummeted; family size is a fraction of what it was 50 years ago; prosperity is on the rise in much of Asia; life expectancy is increasing universally and deaths through wars are in a state of long-term decline.

These developments, good and bad, have a universal character and affect billions of individuals. They involve an understanding of economies as large organisms with interactive components.

Aggregate output in a national economy

A national economy is a complete multi-sector system, made up of household, business, financial, government and international sectors. Each of these sectors is an aggregate or sum of a many smaller economic units with very similar characteristics. The government sector, for example aggregates the taxing and spending activities of local, provincial and national governments. Similarly, the household sector is an aggregate of the income, spending and saving of all households but not the specifics of each individual household. Economic activity within any one of these sectors reflects, in part, the conditions and choices made in that sector. But it is also affects and is affected by conditions and actions in the other sectors. These *interactions* and *feedbacks* within the system mean that the workings of the macro-economy are more complex than the operation of the sum of its parts.

> **Macroeconomics**: the study of the economy as a system in which interactions and feedbacks among sectors determine national output, employment and prices.

For example, consider a simple economy with just household, business and financial sectors. The household sector earns income by providing labour to the other sectors. Households make choices about spending or saving this income. Businesses make decisions about the sizes of their establishments, their labour forces, and their outputs of goods and services. The financial sector provides banking services: bank deposits, loans, and the payments system used by all three sectors.

Suppose households decide to spend more on goods and services and save less. That decision by itself does not change household sector income, but it does increase business sector sales and revenues. It also reduces the flow of household savings into bank deposits in the financial sector. As a result the business sector has an incentive to increase employment and output and perhaps to borrow from the financial sector to finance that expansion. Increased employment in the business sector increases incomes in the household sector and further increases household expenditure and savings. These inter-sector linkages and feedbacks produce a response in aggregate economy greater than the initial change.

Expanding this simple example to include more sectors increases its complexity but does not change the basics. A change in behaviour within a sector,or disturbance from outside that sector, changes aggregate levels of output, employment and prices. A complete multisector macroeconomic theory and model is required to understand the effects, on the aggregate economy, of changes in either internal or external economic conditions. It is also essential for the design of policies to manage the macroeconomic conditions.

Mitigating the effects of a large random shock from outside the economy, like the COVID-19 pandemic, disrupts all sectors of the economy. Flows of income, expenditure, revenue,and output among sectors are reduced sharply both by the pandemic and by government and financial sectors policy responses.

Application Box 1.1: COVID-19 and the Economy

The COVID-19 pandemic attacked Canada in early 2020. It revealed the complexities and interdependencies that drive the macro economy. Control and elimination of the disease depends on stopping person to person transmission. This is why the government mandated personal and social distancing for individuals plus self-isolation and quarantine in some cases. In addition businesses, mainly in the service sector, that relied on face to face interactions with customers or live audiences were forced to close.

As a result, businesses lost sales revenues and cut output. They reduced employment to cut labour costs, but overhead costs remained. Households lost employment and employment incomes. They reduced their discretionary spending, but their overhead costs continued. As a result the economy faced a unique, simultaneous collapse in overall private supply and demand and the risk of a deep recession. The government and financial sectors intervened with fiscal and monetary policy support. Government introduced a wide range of new income supports for the household sector, and loan and subsidy programs to support businesses, funded by large increases in the government's budget deficit. The central bank lowered interest rates and increased the monetary base to support the government's borrowing requirements, and the credit demands on private banks and other financial institutions. The banking system lost the normal growth in customer deposits, but worked to accommodate the needs of their business and household clients.

This unprecedented support from government fiscal policy and central bank monetary policy will offset part of the loss in national output and income. But it will not reverse it. Recovery will begin with the reopening of business and the growth of employment at some time in the uncertain future. The size of the estimated effect of COVID-19 on the Canadian economy is stark. In its Monetary Policy Report, April 2020, The Bank of Canada estimates that real GDP in Canada will be 1% to 7.5% lower in 2020Q1 and lower by 15-30% in 2020Q2 than in 2019Q4. The Monetary Policy Report is available on the Bank of Canada's website at `www.bankofcanada.ca`.

Individual behaviours

Economic actions, at the level of the person or organization, form the subject matter of microeconomics. Formally, **microeconomics** is the study of individual behaviour in the context of scarcity.

Not all individual behaviours are motivated by self-interest; many are motivated by a concern for the well being of society-at-large. Philanthropic societies are goal-oriented and seek to attain their objectives in an efficient manner.

> **Microeconomics**: the study of individual behaviour in the context of scarcity.

Individual economic decisions need not be world-changing events, or motivated by a search for profit. Microeconomics is also about how we choose to spend our time and money. There are quite a few options to choose from: Sleep, work, study, food, shelter, transportation, entertainment, recreation and so forth. Because both time and income are limited we cannot do all things all the time. Many choices are routine or are driven by necessity. You have to eat and you need a place to live. If you have a job you have committed some of your time to work, or if you are a student some of your time is committed to lectures and study. There is more flexibility in other choices. Critically, microeconomics seeks to understand and explain how we make choices and how those choices affect our behaviour in the workplace, the marketplace, and society more generally.

A critical element in making choices is that there exists a *scarcity* of time, or income or productive resources. Decisions are invariably subject to limits or constraints, and it is these constraints that make decisions both challenging and scientific.

Microeconomics also concerns business choices. How does a business use its funds and management skill to produce goods and services? The individual business operator or firm has to decide what to produce, how to produce it, how to sell it and in many cases, how to price it. To make and sell pizza, for example, the pizza parlour needs, in addition to a source of pizza ingredients, a store location (land), a pizza oven (capital), a cook and a sales person (labour). Payments for the use of these inputs generate income to those supplying them. If revenue from the sale of pizzas is greater than the costs of production, the business earns a profit for the owner. A business fails if it cannot cover its costs.

In these micro-level behaviours the decision makers have a common goal: To do as well as they can, *given the constraints imposed by the operating environment*. The individual wants to mix work and leisure in a way that makes her as happy or contented as possible. The entrepreneur aims at making a profit. These actors, or agents as we sometimes call them, are *maximizing*. Such maximizing behaviour is a central theme in this book and in economics at large.

Markets and government

Markets play a key role in coordinating the choices of individuals with the decisions of business. In modern market economies goods and services are supplied by both business and government. Hence we call them **mixed economies**. Some products or services are available through the marketplace to those who wish to buy them and have the necessary income—as in cases like coffee and wireless services. Other services are provided to all people through government programs like law enforcement and health care.

> **Mixed economy**: goods and services are supplied both by private suppliers and government.

Markets offer the choice of a wide range of goods and services at various prices. Individuals can use their incomes to decide the pattern of expenditures and the bundle of goods and services they prefer. Businesses sell goods and services in the expectation that the market price will cover costs and yield a profit.

The market also allows for specialization and separation between production and use. Rather than each individual growing her own food, for example, she can sell her time or labour to employers in return for income. That income can then support her desired purchases. If businesses can produce food more cheaply than individuals the individual obviously gains from using the market – by both having the food to consume, and additional income with which to buy other goods and services. Economics seeks to explain how markets and specialization might yield such gains for individuals and society.

We will represent individuals and firms by envisaging that they have explicit objectives – to maximize their happiness or profit. However, this does not imply that individuals and firms are concerned only with such objectives. On the contrary, much of microeconomics and macroeconomics focuses upon the role of government: How it manages the economy through fiscal and monetary policy, how it redistributes through the tax-transfer system, how it supplies information to buyers and sets safety standards for products.

Since governments perform all of these society-enhancing functions, in large measure governments reflect the social ethos of voters. So, while these voters may be maximizing at the individual level in their everyday lives, and our models of human behaviour in microeconomics certainly emphasize this optimization, economics does not see individuals and corporations as being devoid of civic virtue or compassion, nor does it assume that only market-based activity is important. Governments play a central role in modern economies, to the point where they account for more than one third of all economic activity in the modern mixed economy.

Governments supply goods and services in many spheres, for example, health and education. The provision of public education is motivated both by a concern for equality and a realization that an educated labour force increases the productivity of an economy. Likewise, the provision of law and order, through our legal system broadly defined, represents more than a commitment to a just society at the individual level; without a legal system that enforces contracts and respects property rights, the private sector of the economy would diminish dramatically as a result of corruption, uncertainty and insecurity. It is the lack of such a secure environment in many of the world’s economies that inhibits their growth and prosperity.

Let us consider now the methods of economics, methods that are common to science-based disciplines.

1.2 Understanding through the use of models

Most students have seen an image of Ptolemy's concept of our Universe. Planet Earth forms the centre, with the other planets and our sun revolving around it. The ancients' anthropocentric view of the universe necessarily placed their planet at the centre. Despite being false, this view of our world worked reasonably well – in the sense that the ancients could predict celestial motions, lunar patterns and the seasons quite accurately.

More than one Greek astronomer believed that it was more natural for smaller objects such as the earth to revolve around larger objects such as the sun, and they knew that the sun had to be larger as a result of having studied eclipses of the moon and sun. Nonetheless, the Ptolemaic description of the universe persisted until Copernicus wrote his treatise "On the Revolutions of the Celestial Spheres" in the early sixteenth century. And it was another hundred years before the Church accepted that our corner of the universe is heliocentric. During this time evidence accumulated as a result of the work of Brahe, Kepler and Galileo. The time had come for the Ptolemaic *model* of the universe to be supplanted with a better *model*.

All disciplines progress and develop and explain themselves using models of reality. A **model** is a formalization of theory that facilitates scientific inquiry. Any history or philosophy of science book will describe the essential features of a model. First, it is a stripped down, or reduced, version of the phenomenon that is under study. It incorporates the key elements while disregarding what are considered to be secondary elements. Second, it should accord with reality. Third, it should be able to make meaningful predictions. Ptolemy's model of the known universe met these criteria: It was not excessively complicated (for example distant stars were considered as secondary elements in the universe and were excluded); it corresponded to the known reality of the day, and made pretty good predictions. Evidently not all models are correct and this was the case here.

> **Model**: a formalization of theory that facilitates scientific inquiry.

In short, models are frameworks we use to organize how we think about a problem. Economists sometimes interchange the terms theories and models, though they are conceptually distinct. A **theory** is a logical view of how things work, and is frequently formulated on the basis of observation. A model is a formalization of the essential elements of a theory, and has the characteristics we described above. As an example of an economic model, suppose we theorize that a household's expenditure depends on its key characteristics: A corresponding model might specify that wealth, income, and household size determine its expenditures, while it might ignore other, less important, traits such as the household's neighbourhood or its religious beliefs. The model reduces and simplifies the theory to manageable dimensions. From such a reduced picture of reality we develop an analysis of how an economy and its components work.

> **Theory**: a logical view of how things work, and is frequently formulated on the basis of observation.

An economist uses a model as a tourist uses a map. Any city map misses out some detail—traffic lights and speed bumps, for example. But with careful study you can get a good idea of the best route to take. Economists are not alone in this approach; astronomers, meteorologists, physicists, and genetic scientists operate similarly. Meteorologists disregard weather conditions in South Africa when predicting tomorrow's conditions in Winnipeg. Genetic scientists concentrate on the interactions of limited subsets of genes that they believe are the most important for their purpose. Even with huge computers, all of these scientists build *models* that concentrate on the essentials.

1.3 Opportunity cost and the market

Individuals face choices at every turn: In deciding to go to the hockey game tonight, you may have to forgo a concert; or you will have to forgo some leisure time this week in order to earn additional income for the hockey game ticket. Indeed, there is no such thing as a free lunch, a free hockey game or a free concert. In economics we say that these limits or constraints reflect opportunity cost. The **opportunity cost** of a choice is what must be sacrificed when a choice is made. That cost may be financial; it may be measured in time, or simply the alternative foregone.

Opportunity cost: what must be sacrificed when a choice is made.

Opportunity costs play a determining role in markets. It is precisely because individuals and organizations have different opportunity costs that they enter into exchange agreements. If you are a skilled plumber and an unskilled gardener, while your neighbour is a skilled gardener and an unskilled plumber, then you and your neighbour not only have different capabilities, you also have different opportunity costs, and *you could gain by trading your skills*. Here's why. Fixing a leaking pipe has a low opportunity cost for you in terms of time: You can do it quickly. But pruning your apple trees will be costly because you must first learn how to avoid killing them and this may require many hours. Your neighbour has exactly the same problem, with the tasks in reverse positions. In a sensible world you would fix your own pipes *and* your neighbour's pipes, and she would ensure the health of the apple trees in both backyards.

If you reflect upon this 'sensible' solution—one that involves each of you achieving your objectives while minimizing the time input—you will quickly realize that it resembles the solution provided by the marketplace. You may not have a gardener as a neighbour, so you buy the services of a gardener in the marketplace. Likewise, your immediate neighbour may not need a leaking pipe repaired, but many others in your neighbourhood do, so you sell your service to them. You each specialize in the performance of specific tasks as a result of having different opportunity costs or different efficiencies. Let us now develop a model of exchange to illustrate the advantages of specialization and trade, and hence the markets that facilitate these activities. This model is developed with the help of some two-dimensional graphics.

1.4 A model of exchange and specialization

Production and specialization

We have two producers and two goods: Amanda and Zoe produce vegetables (V) and or fish (F). Their production capabilities are defined in Table 1.1 and in Figure 1.1, where the quantity of V appears on the vertical axis and the quantity of F on the horizontal axis. Zoe and Amanda each have 36-hour weeks and they devote that time to producing the two goods. But their efficiencies differ: Amanda requires two hours to produce a unit of V and three hours for a unit of F. As a consequence, if she devotes all of her time to V she can produce 18 units, or if she devotes all of her time to F she can produce 12 units. Or, she could share her time between the two. This environment can also be illustrated and analyzed graphically, as in Figure 1.1.

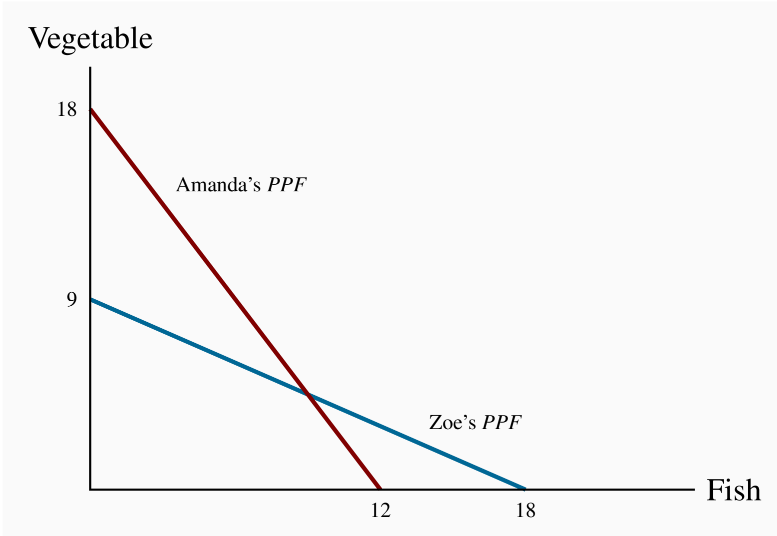
Table 1.1: Production possibilities in a two-person economy

	Hours/ fish	Hours/ vegetable	Fish specialization	Vegetable specialization
Amanda	3	2	12	18
Zoe	2	4	18	9

Each producer has a time allocation of 36 hours. By allocating total time to one activity, Amanda can produce $12F$ or $18V$, Zoe can produce $18F$ or $9V$. By splitting their time each person can also produce a combination of the two.

Two-dimensional graphics are a means of portraying the operation of a model, as defined above. We will use these graphical representations throughout the text. In this case, Amanda's production capability is represented by the line that meets the vertical axis at 18 and the horizontal axis at 12. The vertical point indicates that she can produce 18 units of V if she produces zero units of F – keep in mind that where V has a value of 18, Amanda has no time left for fish production. Likewise, if she devotes all of her time to fish she can produce 12 units, since each unit requires 3 of her 36 hours. The point $F = 12$ is thus another possibility for her. In addition to these two possibilities, which we can term 'specialization', she could allocate her time to producing some of each good. For example, by dividing her 36 hours equally she could produce 6 units of F and 9 units of V. A little computation will quickly convince us that different allocations of her time will lead to combinations of the two goods that lie along a straight line joining the specialization points.

Figure 1.1: Absolute advantage – production



Amanda's *PPF* indicates that she can produce either $18V$ (and zero F), or $12F$ (and zero V), or some combination. Zoe's *PPF* indicates she can produce either $9V$ (and zero F), or $18F$ (and zero V), or some combination. Amanda is more efficient in producing V and Zoe is more efficient at producing F.

We will call this straight line Amanda's **production possibility frontier (*PPF*)**: It is the combination of goods she can produce while using all of her resources – time. She could not produce combinations of goods represented by points beyond this line (to the top right). She could indeed produce combinations below it (lower left) – for example, a combination of 4 units of V and 4 units of F; but such points would not require all of her time. The $(4,4)$ combination would require just 20 hours. In sum, points beyond this line are not feasible, and points within it do not require all of her time resources.

> **Production possibility frontier (*PPF*)**: the combination of goods that can be produced using all of the resources available.

Having developed Amanda's *PPF*, it is straightforward to develop a corresponding set of possibilities for Zoe. If she requires 4 hours to produce a unit of V and 2 hours to produce a unit of F, then her 36 hours will enable her to specialize in 9 units of V or 18 units of F; or she could produce a combination represented by the straight line that joins these two specialty extremes.

Consider now the opportunity costs for each person. Suppose Amanda is currently producing 18 V and zero F, and considers producing some F and less V. For each unit of F she wishes to produce, it is evident from her *PPF* that she must sacrifice 1.5 units of V. This is because F requires 50% more hours than V. Her trade-off is 1.5 : 1.0. The additional time requirement is also expressed in the intercept values: She could give up 18 units of V and produce 12 units of F instead; this again is a ratio of 1.5 : 1.0. This ratio defines her opportunity cost: The cost of an additional unit of F is that 1.5 units of V must be 'sacrificed'.

Applying the same reasoning to Zoe's *PPF*, her opportunity cost is 0.5 : 1; she must sacrifice one half of a unit of V to free up enough time to produce one unit of F.

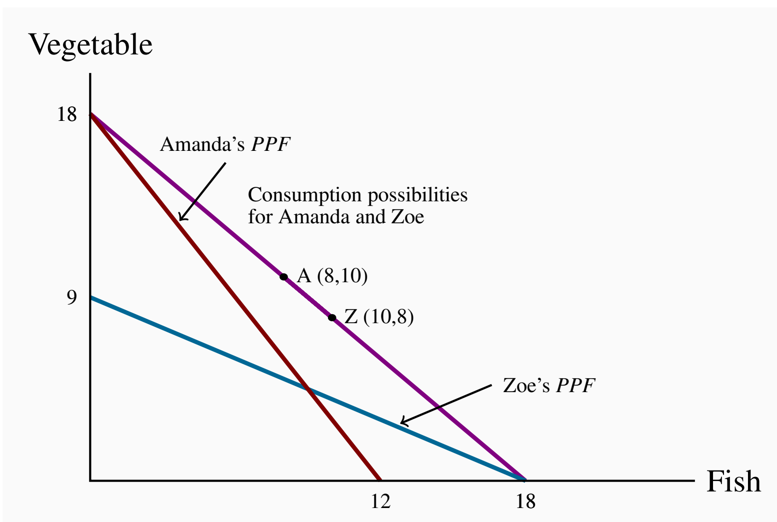
So we have established two things about Amanda and Zoe's production possibilities. First, if Amanda specializes in V she can produce more than Zoe, just as Zoe can produce more than Amanda if Zoe specializes in F. Second, their opportunity costs are different: Amanda must sacrifice more V than Zoe in producing one more unit of F. The different opportunity costs translate into potential gains for each individual.

The gains from exchange

We shall illustrate the gains that arise from specialization and exchange graphically. Note first that if these individuals are self-sufficient, in the sense that they consume their own production, each individual's consumption combination will lie on their own *PPF*. For example, Amanda could allocate half of her time to each good, and produce (and consume) $6F$ and $9V$. Such a point necessarily lies on her *PPF*. Likewise for Zoe. So, *in the absence of exchange*, each individual's *PPF* is also her **consumption possibility frontier** (*CPF*). In Figure 1.1 the *PPF* for each individual is thus also her *CPF*.

> **Consumption possibility frontier** (*CPF*): the combination of goods that can be consumed as a result of a given production choice.

Figure 1.2: Absolute advantage – consumption



With specialization and trade at a rate of 1 : 1 they consume along the line joining the specialization points. If Amanda trades $8V$ to Zoe in return for $8F$, Amanda moves to the point A(8,10) and Zoe to Z(10,8). Each can consume more after specialization than before specialization.

Upon realizing that they are not equally efficient in producing the two goods, they decide to specialize completely in producing just the single good where they are most efficient. Amanda specializes in V and Zoe in F. Next they must agree to a rate at which to exchange V for F. Since Amanda's opportunity cost is 1.5 : 1 and Zoe's is 0.5 : 1, suppose they agree to exchange V for F at an intermediate rate of 1 : 1. There are many trading, or exchange, rates possible; our purpose is to illustrate that gains are possible for *both* individuals at some exchange rate. The choice of this rate also makes the graphic as simple as possible. At this exchange rate, $18V$ must exchange for $18F$. In Figure 1.2, this means that each individual is now able to consume along the line joining the coordinates $(0,18)$ and $(18,0)$.[1] This is because Amanda produces $18V$ and she can trade at a rate of 1 : 1, while Zoe produces $18F$ and trades at the same rate of 1 : 1.

The fundamental result illustrated in Figure 1.2 is that, as a result of specialization and trade, each individual can consume combinations of goods that lie on a line beyond her initial consumption possibilities. Their consumption well-being has thus improved. For example, suppose Amanda trades away $8V$ to Zoe and obtains $8F$ in return. The points 'A' and 'Z' with coordinates $(8,10)$ and $(10,8)$ respectively define their final consumption. Pre-specialization, if Amanda wished to consume $8F$ she would have been constrained to consume $6V$ rather than the $10V$ now possible. Zoe benefits correspondingly.[2]

The foregoing example illustrates that trade is not a **zero-sum game**; it has a positive net value because both parties to the trade can gain. A zero-sum gain is where the gains to one party exactly offset the losses to another. This is an extraordinarily important principle in trade negotiations, whether international or domestic.

> **A zero-sum game** is an interaction where the gain to one party equals the loss to another party.

Market design

In the preceding example we have shown that specialization provides scope for gains that can accrue to those participating in the exchange. But this tells us little about how a market for these products comes into being: how does the exchange take place, and how is information transmitted? The answer is that while some markets have evolved historically to their current state, many markets are designed by an institution or a firm. Fruit and vegetable markets have been with us for thousands of years - since we ceased being purely a hunter-gatherer society. They exist in every community in the world economy. In contrast, the Dutch tulip auction was designed in the early 1600s and exists in basically the same form to this day: the auctioneer begins with a high price, lowers it at known time intervals (measured in seconds or minutes) until some buyer signals that

[1] When two values, separated by a comma, appear in parentheses, the first value refers to the horizontal-axis variable, and the second to the vertical-axis variable.

[2] In the situation we describe above one individual is absolutely more efficient in producing one of the goods and absolutely less efficient in the other. We will return to this model in Chapter 15 and illustrate that consumption gains of the type that arise here can also result if one of the individuals is absolutely more efficient in producing both goods, but that the degree of such advantage differs across goods.

she is willing to purchase the lot on offer. Supermarkets in contrast offer goods at a fixed price. Government contracts are normally signed after a tendering process, in which interested suppliers submit bids. Amazon Inc. is currently experimenting with cashierless 'bricks and mortar' stores that monitor all transactions electronically. Craig's List and E-Bay have their own sets of rules.

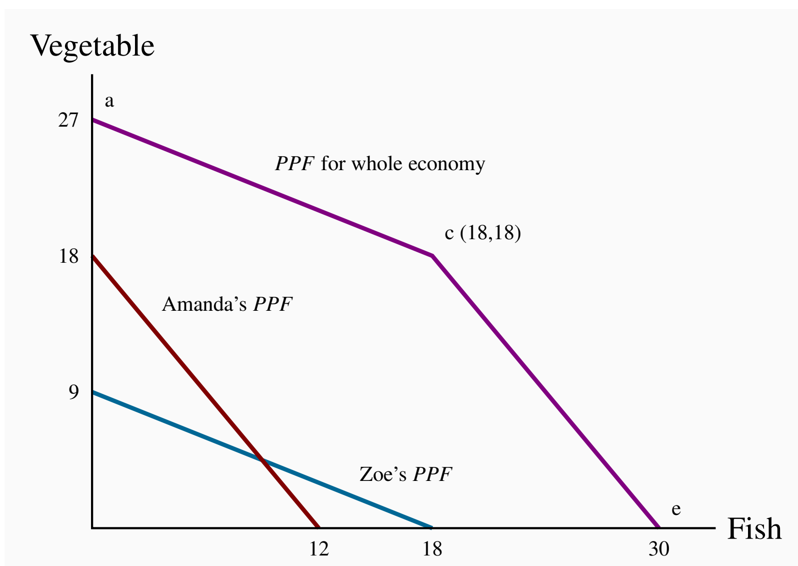
In each of these cases markets are designed, frequently with a specific objective on the part of the supplier or the mediating institution: Amazon wants to increase its share of all goods trades; governments wish to limit costs. Markets do not all grow spontaneously and the structure of a market will influence how the gains from trade are distributed.

1.5 Economy-wide production possibilities

The *PPF*s in Figures 1.1 and 1.2 define the amounts of the goods that each *individual* can produce while using all of their fixed productive capacity—time in this instance. The national, or economy-wide, *PPF* for this two-person economy reflects these individual possibilities combined. Such a frontier can be constructed using the individual frontiers as the component blocks.

First let us define this economy-wide frontier precisely. The **economy-wide** *PPF* is the set of goods and services combinations that can be produced in the economy when all available productive resources are in use. Figure 1.3 contains both of the individual frontiers plus the aggregate of these, represented by the kinked line *ace*. The point on the V axis, $a = 27$, represents the total amount of V that could be produced if both individuals devoted all of their time to it. The point $e = 30$ on the horizontal axis is the corresponding total for fish.

Figure 1.3: Economy-wide PPF



From a, to produce Fish it is more efficient to use Zoe because her opportunity cost is less (segment ac). When Zoe is completely specialized, Amanda produces (ce). With complete specialization this economy can produce $27V$ or $30F$.

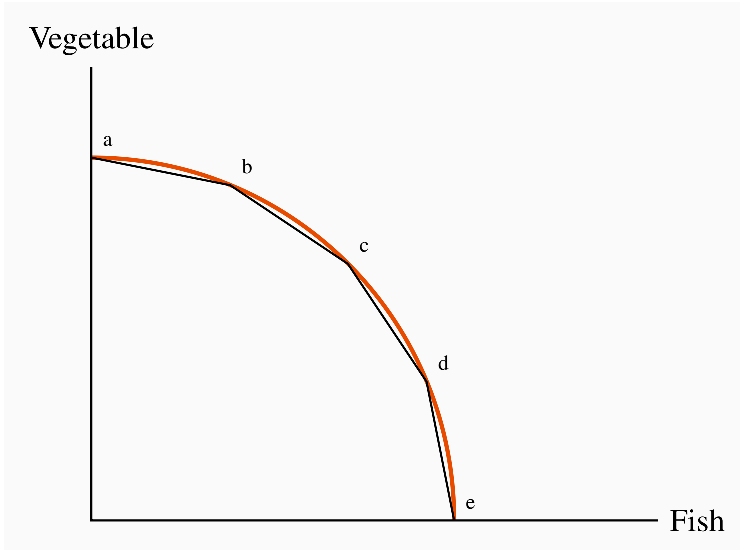
> **Economy-wide** *PPF*: the set of goods and services combinations that can be produced in the economy when all available productive resources are in use.

To understand the point c, imagine initially that all resources are devoted to V. From such a point, a, consider a reduction in V and an increase in F. The most efficient way of increasing F production at the point a is to use the individual whose opportunity cost is lower. Zoe can produce one unit of F by sacrificing just 0.5 units of V, whereas Amanda must sacrifice 1.5 units of V to produce 1 unit of F. Hence, at this stage Amanda should stick to V and Zoe should devote some time to fish. In fact as long as we want to produce more fish Zoe should be the one to do it, until she has exhausted her time resource. This occurs after she has produced $18F$ and has ceased producing V. At this point the economy will be producing $18V$ and $18F$ – the point c.

From this combination, if the economy wishes to produce more fish Amanda must become involved. Since her opportunity cost is 1.5 units of V for each unit of F, the next segment of the economy-wide PPF must see a reduction of 1.5 units of V for each additional unit of F. This is reflected in the segment ce. When both producers allocate all of their time to F the economy can produce 30 units. Hence the economy's PPF is the two-segment line ace. Since this has an outward kink, we call it concave (rather than convex).

As a final step consider what this PPF would resemble if the economy were composed of many persons with differing efficiencies. A little imagination suggests (correctly) that it will have a segment for each individual and continue to have its outward concave form. Hence, a four-person economy in which each person had a different opportunity cost could be represented by the segmented line $abcde$, in Figure 1.4. Furthermore, we could represent the PPF of an economy with a very large number of such individuals by a somewhat smooth PPF that accompanies the 4-person PPF. The logic for its shape continues to be the same: As we produce less V and more F we progressively bring into play resources, or individuals, whose opportunity cost, in terms of reduced V is higher.

Figure 1.4: A multi-person PPF



The PPF for the whole economy, abcde, is obtained by allocating productive resources most efficiently. With many individuals we can think of the PPF as the *concave envelope* of the individual capabilities.

The outputs V and F in our economic model require just one input – time, but if other productive resources were required the result would be still a concave PPF. Furthermore, we generally interpret the PPF to define the output possibilities *when the economy is running at its normal capacity*. In this example, we consider a work week of 36 hours to be the 'norm'. Yet it is still possible that the economy's producers might work some additional time in exceptional circumstances, and this would increase total production possibilities. This event would be represented by an outward movement of the PPF.

1.6 Aggregate output, growth and business cycles

The PPF can be used to illustrate several aspects of macroeconomics: In particular, the level of an economy's output, the growth of national and per capita output over time, and short-run business-cycle fluctuations in national output and employment.

Aggregate output

An economy's capacity to produce goods and services depends on its endowment of resources and the productivity of those resources. The two-person, two-product examples in the previous section reflect this.

The **productivity of labour**, defined as output per worker or per hour, depends on:

- The skill, knowledge and experience of the labour force;

- The **capital stock**: Buildings, machinery, equipment, and software the labour force has to work with; and
- The current state of technology.

> The **productivity of labour** is the output of goods and services per worker.
>
> An economy's **capital stock** is the buildings, machinery, equipment and software used in producing goods and services.

The economy's output, which we define by Y, can be defined as the output per worker times the number of workers; hence, we can write:

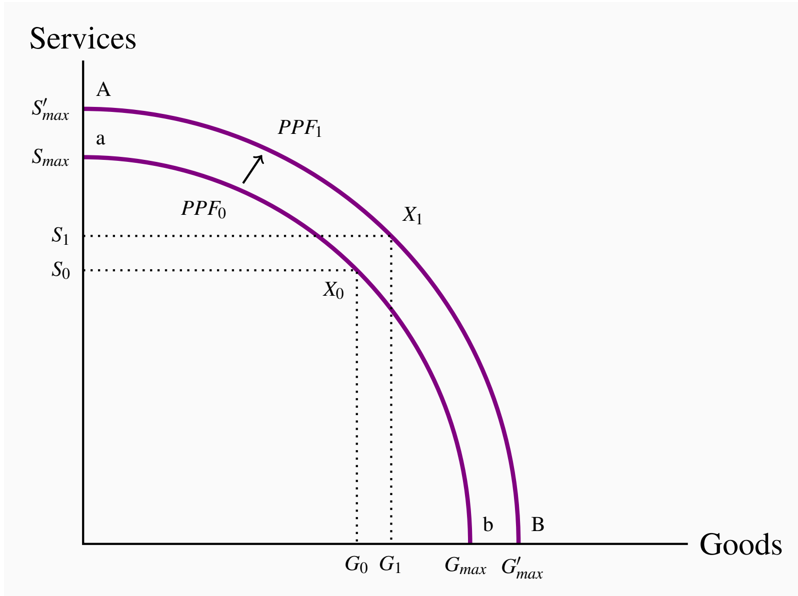
$$Y = (\text{number of workers employed}) \times (\text{output per worker}).$$

When the employment of labour corresponds to 'full employment' in the sense that everyone willing to work at current wage rates and normal hours of work is working, the economy's actual output is also its capacity output Y_c. We also term this capacity output as **full employment output**:

> **Full employment output** $Y_c = (\text{number of workers at full employment}) \times (\text{output per worker})$.

Suppose the economy is operating with full employment of resources producing outputs of two types: Goods and services. In Figure 1.5, PPF_0 shows the different combinations of goods and services the economy can produce in a particular year using all its labour, capital and the best technology available at the time.

Figure 1.5: Growth and the PPF



Economic growth is illustrated by an outward shift in the *PPF* from PPF_0 to PPF_1. PPF_1 shows the economy can produce more in both sectors than with PPF_0.

An aggregate economy produces a large variety of outputs in two broad categories. Goods are the products of the agriculture, forestry, mining, manufacturing and construction industries. Services are provided by the wholesale and retail trade, transportation, hospitality, finance, health care, education, legal and other service sectors. As in the two-product examples used earlier, the shape of the *PPF* illustrates the opportunity cost of increasing the output of either product type. We are not concerned with who supplies the products for the moment: It may be the private sector or the government.

Point X_0 on PPF_0 shows one possible structure of capacity output. This combination may reflect the pattern of demand and hence expenditures in this economy. Output structures and therefore the shapes of *PPF*s differ among economies with different income levels. High-income economies spend more on services than goods and produce higher ratios of services to goods. Middle income countries produce lower ratios of services to goods, and low income countries much lower ratios of services to goods. For example, in 2017, the structure of national output in Canada was 70 percent services and 30 percent goods, while in Mexico the structure was 48 percent services and 52 percent goods.

Different countries also have different *PPF*s and different output structures, depending on their resource endowments, labour forces, capital stocks, technology and expenditure patterns.

Economic growth

Three things contribute to growth in the economy. The labour supply grows as the population expands; the stock of capital grows as spending by business (and government) on buildings, machinery, information technology and so forth increases; and labour-force productivity grows as a result of experience, the development of scientific knowledge combined with product and process innovations, and advances in the technology of production. Combined, these developments expand capacity output over time. In Figure 1.5 economic growth shifts the PPF out from PPF_0 to PPF_1.

This basic description covers the key sources of growth in total output. Economies differ in their rates of overall economic growth as a result of different rates of growth in labour force, in capital stock, and improvements in technology. But improvements in standards of living require more than growth in total output. Increases in output *per worker* and *per person* are necessary. Sustained increases in living standards require sustained growth in labour productivity, which in turn is based on advances in the technology along with the amount of capital each worker has to work with. Furthermore, if the growth in output is to benefit society at large, workers across the board need to see an increase in their earnings. As we shall explore in Chapter 13, several developed countries have seen the fruits of growth concentrated in the hands of the highest income earners.

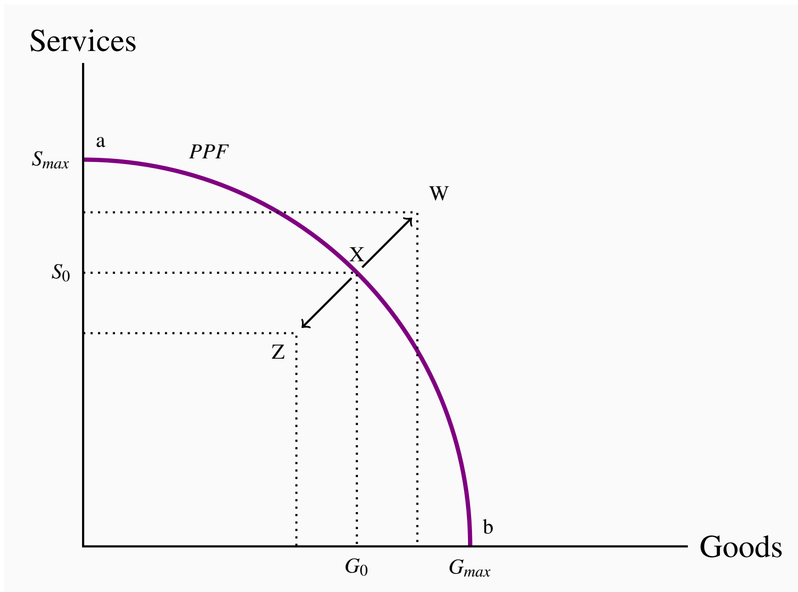
Recessions and booms

A prime objective of economic policy is to ensure that the economy operates on or near the PPF – it should use its resources to capacity and have minimal unemployment. However, economic conditions are seldom tranquil for long periods of time. Unpredictable changes in business expectations of future profits, in consumer confidence, in financial markets, in commodity and energy prices, in trade agreements and disputes, in economic conditions in major trading partners, in government policy and many other events disrupt patterns of expenditure and output. Some of these changes disturb the level of total expenditure and thus the demand for total output. Others disturb the conditions of production and thus the economy's production capacity. Whatever the exact cause, the economy may be pushed off its current PPF. If expenditures on goods and services decline, the economy may experience a **recession**. Output would fall short of capacity output and unemployment would rise. Alternatively, times of rapidly growing expenditure and output may result in an economic **boom**: Output and employment expand beyond capacity levels.

> An **economic recession** occurs when output falls below the economy's capacity output.
>
> A **boom** is a period of high growth that raises output above normal capacity output.

Recent history provides examples. Following the financial crisis of 2008-09 that hit the US and many other developed economies, many economies were pushed into recessions. Expenditure on new residential construction collapsed for lack of income and secure financing, as did business investment, consumption spending and exports. Lower expenditures reduced producers' revenues, forcing cuts in output and employment and reducing household incomes. Lower incomes led to further cutbacks in spending. In Canada in 2009 aggregate output declined by 2.9 percent, employment declined by 1.6 percent and the unemployment rate rose from 6.1 percent in 2008 to 8.3 percent by 2010. The world's economies have been slow to recover, and even by 2019 the output in several developed economies was no higher than it was in 2008. Canada's recession was not nearly as severe as the recessions in economies such as Spain, Italy and Greece; but output between 2009 and 2019 has been below the potential of the Canadian economy. In the third quarter of 2019 the national output was about 0.7 percent below potential output and the unemployment rate was 5.5 percent.

Figure 1.6: Booms and recessions



Economic recessions leave the economy below its normal capacity; the economy might be driven to a point such as Z. Economic expansions, or booms, may drive capacity above its normal level, to a point such as W.

An economy in a recession is operating inside its PPF. The fall in output from X to Z in Figure 1.6 illustrates the effect of a recession. Expenditures on goods and services have declined. Output is less than capacity output, unemployment is up and some plant capacity is idle. Labour income and business profits are lower. More people would like to work and business would like to produce and sell more output, but it takes time for interdependent product, labour and financial markets in the economy to adjust and increase employment and output. Monetary and fiscal policy may be productive in specific circumstances, to stimulate demand, increase output and employment and move the economy back to capacity output and full employment. The development and implementation of such policies form the core of macroeconomics.

Alternatively, an unexpected increase in demand for exports would increase output and employment. Higher employment and output would increase incomes and expenditure, and in the process spread the effects of higher output sales to other sectors of the economy. The economy would move outside its PPF, for example to W in Figure 1.6, by using its resources more intensively than normal. Unemployment would fall and overtime work would increase. Extra production shifts would run plant and equipment for longer hours and work days than were planned when it was designed and installed. Output at this level may not be sustainable, because shortages of labour and materials along with excessive rates of equipment wear and tear would push costs and prices up. Again, we will examine how the economy reacts to such a state in our macroeconomic analysis.

Output and employment in the Canadian economy over the past twenty years fluctuated about growth trend in the way Figure 1.6 illustrates. For several years prior to 2008 the Canadian economy operated slightly above its capacity; but once the recession arrived monetary and fiscal policy were used to fight it – to bring the economy back from a point such as Z towards a point such as X on the PPF.

Macroeconomic models and policy

The PPF diagrams illustrate the main dimensions of macroeconomics: Capacity output, growth in capacity output and business cycle fluctuations in actual output relative to capacity. But these diagrams do not offer explanations and analysis of macroeconomic activity. We need a macroeconomic *model* to understand and evaluate the causes and consequences of business cycle fluctuations. As we shall see, these models are based on explanations of expenditure decisions by households and business, financial market conditions, production costs and producer pricing decisions at different levels of output. Models also capture the objectives of fiscal and monetary policies and provide a framework for policy evaluation. A full macroeconomic model integrates different sector behaviours and the feedbacks across sectors that can moderate or amplify the effects of changes in one sector on national output and employment.

Conclusion

We have covered a lot of ground in this introductory chapter. It is intended to open up the vista of economics to the new student in the discipline. Economics is powerful and challenging, and the ideas we have developed here will serve as conceptual foundations for our exploration of the subject.

Key Terms

Macroeconomics studies the economy as system in which linkages and feedbacks among sectors determine national output, employment and prices.

Microeconomics is the study of individual behaviour in the context of scarcity.

Mixed economy: goods and services are supplied both by private suppliers and government.

Model is a formalization of theory that facilitates scientific inquiry.

Theory is a logical view of how things work, and is frequently formulated on the basis of observation.

Opportunity cost of a choice is what must be sacrificed when a choice is made.

Production possibility frontier (*PPF*) defines the combination of goods that can be produced using all of the resources available.

Consumption possibility frontier (*CPF*): the combination of goods that can be consumed as a result of a given production choice.

A **zero-sum game** is an interaction where the gain to one party equals the loss to another party.

Economy-wide *PPF* is the set of goods combinations that can be produced in the economy when all available productive resources are in use.

Productivity of labour is the output of goods and services per worker.

Capital stock: the buildings, machinery, equipment and software used in producing goods and services.

Full employment output $Y_c =$ (number of workers at full employment) $\times$ (output per worker).

Recession: when output falls below the economy's capacity output.

Boom: a period of high growth that raises output above normal capacity output.

Exercises for Chapter 1

Exercise 1.1 An economy has 100 identical workers. Each one can produce four cakes or three shirts, regardless of the number of other individuals producing each good.

(a) How many cakes can be produced in this economy when all the workers are cooking?

(b) How many shirts can be produced in this economy when all the workers are sewing?

(c) On a diagram with cakes on the vertical axis, and shirts on the horizontal axis, join these points with a straight line to form the PPF.

(d) Label the inefficient and unattainable regions on the diagram.

Exercise 1.2 In the table below are listed a series of points that define an economy's production possibility frontier for goods Y and X.

Y	1000	900	800	700	600	500	400	300	200	100	0
X	0	1600	2500	3300	4000	4600	5100	5500	5750	5900	6000

(a) Plot these pairs of points to scale, on graph paper, or with the help of a spreadsheet.

(b) Given the shape of this PPF is the economy made up of individuals who are similar or different in their production capabilities?

(c) What is the opportunity cost of producing 100 more Y at the combination $(X = 5500, Y = 300)$.

(d) Suppose next there is technological change so that at every output level of good Y the economy can produce 20 percent more X. Enter a new row in the table containing the new values, and plot the new PPF.

Exercise 1.3 Using the PPF that you have graphed using the data in Exercise 1.2, determine if the following combinations are attainable or not: $(X = 3000, Y = 720)$, $(X = 4800, Y = 480)$.

Exercise 1.4 You and your partner are highly efficient people. You can earn \$20 per hour in the workplace; your partner can earn \$30 per hour.

(a) What is the opportunity cost of one hour of leisure for you?

(b) What is the opportunity cost of one hour of leisure for your partner?

(c) Now consider what a PPF would look like: You can produce/consume two things, leisure and income. Since income buys things you can think of the PPF as having these two 'products' – leisure and consumption goods/services. So, with leisure on the horizontal axis and income in dollars is on the vertical axis, plot your PPF. You can assume that you have 12 hours per day to allocate to either leisure or income. [*Hint*: the leisure axis will have an intercept of 12 hours. The income intercept will have a dollar value corresponding to where all hours are devoted to work.]

(d) Draw the *PPF* for your partner.

Exercise 1.5 Louis and Carrie Anne are students who have set up a summer business in their neighbourhood. They cut lawns and clean cars. Louis is particularly efficient at cutting the grass – he requires one hour to cut a typical lawn, while Carrie Anne needs one and one half hours. In contrast, Carrie Anne can wash a car in a half hour, while Louis requires three quarters of an hour.

(a) If they decide to specialize in the tasks, who should cut the grass and who should wash cars?

(b) If they each work a twelve hour day, how many lawns can they cut and how many cars can they wash if they each specialize in performing the task where they are most efficient?

(c) Illustrate the *PPF* for each individual where lawns are on the horizontal axis and car washes on the vertical axis, if each individual has twelve hours in a day.

Exercise 1.6 Continuing with the same data set, suppose Carrie Anne's productivity improves so that she can now cut grass as efficiently as Louis; that is, she can cut grass in one hour, and can still wash a car in one half of an hour.

(a) In a new diagram draw the *PPF* for each individual.

(b) In this case does specialization matter if they are to be as productive as possible as a team?

(c) Draw the *PPF* for the whole economy, labelling the intercepts and the 'kink' point coordinates.

Exercise 1.7 Going back to the simple *PPF* plotted for Exercise 1.1 where each of 100 workers can produce either four cakes or three shirts, suppose a recession reduces demand for the outputs to 220 cakes and 129 shirts.

(a) Plot this combination of outputs in the diagram that also shows the *PPF*.

(b) How many workers are needed to produce this output of cakes and shirts?

(c) What percentage of the 100 worker labour force is unemployed?

Chapter 2

Theories, data and beliefs

In this chapter we will explore:

Economists, like other scientists and social scientists, observe and analyze behaviour and events. Economists are concerned primarily with the economic causes and consequences of what they observe. They want to understand an extensive range of human experience, including: money, government finances, industrial production, household consumption, inequality in income distribution, war, monopoly power, health, professional and amateur sports, pollution, marriage, the arts, and much more.

Economists approach these issues using theories and models. To present, explain, illustrate and evaluate their theories and models they have developed a set of techniques or tools. These involve verbal descriptions and explanations, diagrams, algebraic equations, data tables and charts and statistical tests of economic relationships.

This chapter covers some of these basic techniques of analysis.

2.1 Data analysis

The analysis of behaviour necessarily involves data. Data may serve to validate or contradict a theory. Data analysis, even without being motivated by economic theory, frequently displays patterns of behaviour that merit examination. The terms *variables* and *data* are related. **Variables** are measures that can take on different magnitudes. The interest rate on a student loan, for example, is a variable with a certain value at a point in time but perhaps a different value at an earlier or later date. Economic theories and models explain the causal relationships between variables. In contrast, **Data** are the recorded values of variables. Sets of data provide specific values for the variables we want to study and analyze. Knowing that gross domestic product (a variable) declined in 2009 is just a partial description of events. If the data indicate that it decreased by exactly 3%, we know a great deal more – the decline was large.

> **Variables**: measures that can take on different values.

Data: recorded values of variables.

Sets of data help us to test our models or theories, but first we need to pay attention to the economic logic involved in observations and modelling. For example, if sunspots or baggy pants were found to be correlated with economic expansion, would we consider these events a coincidence or a key to understanding economic growth? The observation is based on facts or data, but it need not have any economic content. The economist's task is to distinguish between coincidence and economic causation. Merely because variables are associated or correlated does not mean that one causes the other.

While the more frequent wearing of loose clothing in the past may have been associated with economic growth because they both occurred at the same time (correlation), one could not argue on a logical basis that this behaviour causes good economic times. Therefore, the past association of these variables should be considered as no more than a coincidence. Once specified on the basis of economic logic, a model must be tested to determine its usefulness in explaining observed economic events.

Table 2.1: House prices and price indexes

Year	House prices in dollars (P_H)	Percentage change in P_H	Percentage change in consumer prices	Real percentage change in P_H	Index for price of housing	5-year mortgage rate
2001	350,000				100	7.75
2002	360,000				102.9	6.85
2003	395,000	35,000/360,000=9.7%	3%	6.7%	112.9	6.6
2004	434,000				124.0	5.8
2005	477,000				136.3	6.1
2006	580,000				165.7	6.3
2007	630,000				180.0	6.65
2008	710,000				202.9	7.3
2009	605,000	-105,000/710,000=-14.8%	1.6%	-16.4%	172.9	5.8
2010	740,000				211.4	5.4
2011	800,000				228.6	5.2

Note: Data on changes in consumer prices come from Statistics Canada, CANSIM series V41692930; data on house prices are for N. Vancouver from *Royal Le Page*; data on mortgage rates from `http://www.ratehub.ca`. Index for house prices obtained by scaling each entry in column 2 by 100/350,000. The real percentage change in the price of housing is: The percentage change in the price of housing minus the percentage change in consumer prices.

Data types

Data come in several forms. One form is **time-series**, which reflects a set of measurements made in sequence at different points in time. The first column in Table 2.1 reports the values for house prices in North Vancouver for the first quarter of each year, between 2001 and 2011. Evidently this is a time series. Annual data report one observation per year. We could, alternatively, have presented the data in monthly, weekly, or even daily form. The frequency we use depends on the purpose: If we are interested in the longer-term trend in house prices, then the annual form suffices. In contrast, financial economists, who study the behaviour of stock prices, might not be content with daily or even hourly prices; they may need prices minute-by-minute. Such data are called **high-frequency** data, whereas annual data are **low-frequency** data.

Table 2.2: Unemployment rates, Canada and Provinces, monthly 2012, seasonally adjusted

	Jan	Feb	Mar	Apr	May	Jun
CANADA	7.6	7.4	7.2	7.3	7.3	7.2
NFLD	13.5	12.9	13.0	12.3	12.0	13.0
PEI	12.2	10.5	11.3	11.0	11.3	11.3
NS	8.4	8.2	8.3	9.0	9.2	9.6
NB	9.5	10.1	12.2	9.8	9.4	9.5
QUE	8.4	8.4	7.9	8.0	7.8	7.7
ONT	8.1	7.6	7.4	7.8	7.8	7.8
MAN	5.4	5.6	5.3	5.3	5.1	5.2
SASK	5.0	5.0	4.8	4.9	4.5	4.9
ALTA	4.9	5.0	5.3	4.9	4.5	4.6
BC	6.9	6.9	7.0	6.2	7.4	6.6

Source: Statistics Canada CANSIM Table 282-0087.

> **Time-series**: a set of measurements made sequentially at different points in time.
>
> **High (low) frequency data**: series with short (long) intervals between observations.

In contrast to time-series data, **cross-section** data record the values of different variables at a point in time. Table 2.2 contains a cross-section of unemployment rates for Canada and Canadian provinces economies. For January 2012 we have a snapshot of the provincial economies at that point in time, likewise for the months until June. This table therefore contains **repeated cross-sections**.

When the unit of observation is the same over time such repeated cross sections are called longitudinal data. For example, a health survey that followed and interviewed the same individuals over time would yield longitudinal data. If the individuals differ each time the survey is conducted, the data are repeated cross sections. **Longitudinal data** therefore follow the same units of observation through time.

> **Cross-section data**: values for different variables recorded at a point in time.
>
> **Repeated cross-section data**: cross-section data recorded at regular or irregular intervals.
>
> **Longitudinal data**: follow the same units of observation through time.

Graphing the data

Data can be presented in graphical as well as tabular form. Figure 2.1 plots the house price data from the second column of Table 2.1. Each asterisk in the figure represents a price value and a corresponding time period. The horizontal axis reflects time, the vertical axis price in dollars. The graphical presentation of data simply provides a visual rather than numeric perspective. It is immediately evident that house prices increased consistently during this 11-year period, with a single downward 'correction' in 2009. We have plotted the data a second time in Figure 2.2 to illustrate the need to read graphs carefully. The greater *apparent* slope in Figure 2.1 might easily be interpreted to mean that prices increased more steeply than suggested in Figure 2.2. But a careful reading of the axes reveals that this is not so; using different scales when plotting data or constructing diagrams can mislead the unaware viewer.

Figure 2.1: House prices in dollars 1999-2012

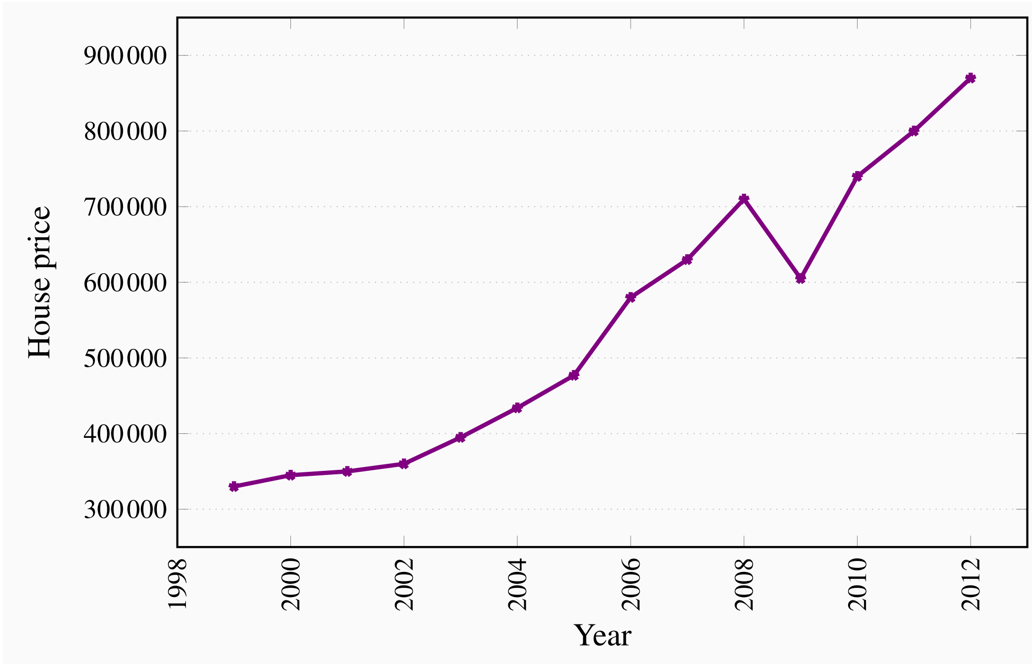
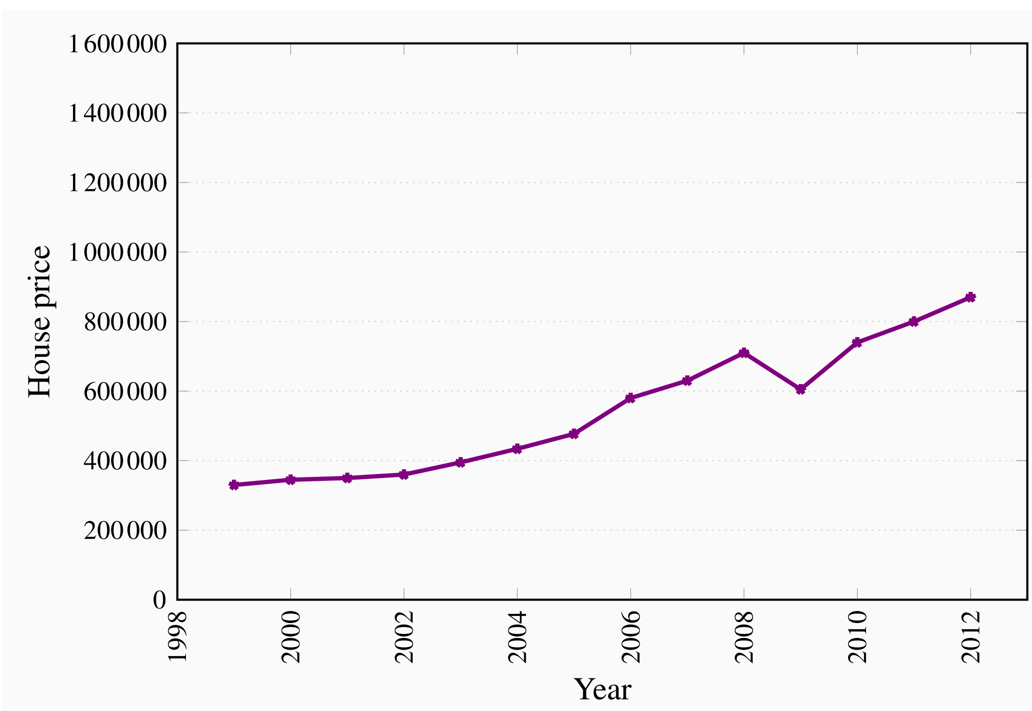


Figure 2.2: House prices in dollars 1999-2012



Percentage changes

The use of percentages makes the analysis of data particularly simple. Suppose we wanted to compare the prices of New York luxury condominiums with the prices of homes in rural Mississippi. In the latter case, a change in average prices of $10,000 might be considered enormous, whereas a change of one million dollars in New York might be pretty normal – because the average price in New York is so much higher than in Mississippi. To make comparisons between the two markets, we can use the concept of a **percentage change**. This is defined as the change in the value of the variable, relative to its initial value, multiplied by 100.

> **Percentage change**= [(change in values)/(original value)] × 100.

The third column of Table 2.1 contains the values of the percentage change in house prices for two pairs of years. Between 2002 and 2003 the price change was $35,000. Relative to the price in the first of these two years this change was the fraction $35,000/395,000 = 0.097$. If we multiply this fraction by 100 we obtain a percentage price change of 9.7%. Evidently we could calculate the percentage price changes for all pairs of years. A second price change is calculated for the 2008-2009 pair of years. Here price declined and the result is thus a negative percentage change.

Consumer prices

Most variables in economics are averages of the components that go into them. When variables are denominated in dollar terms it is important to be able to interpret them correctly. While the house price series above indicates a strong pattern of price increases, it is vital to know if the price of housing increased more or less rapidly that other prices in the economy. If all prices in the economy were increasing in line with house prices there would be no special information in the house price series. However, if house prices increased more rapidly than prices in general, then the data indicate that something special took place in the housing market during the decade in question. To determine an answer to this we need to know the degree to which the general price level changed each year.

Statistics Canada regularly surveys the price of virtually every product produced in the economy. One such survey records the prices of goods and services purchased by consumers. Statistics Canada then computes an average price level for all of these goods combined for each time period the survey is carried out (monthly). Once Statistics Canada has computed the average consumer price, it can compute the change in the price level from one period to the next. In Table 2.1 two such values are entered in the following data column: Consumer prices increased by 3% between 2002 and 2003, and by 1.6% between 2008 and 2009. These percentage changes in the general price level represent **inflation** if prices increase, and **deflation** if prices decline.

In this market it is clear that housing price changes were substantially larger than the changes in consumer prices for these two pairs of years. The next column provides information on the difference between the house price changes and changes in the general consumer price level, in percentage terms. This is (approximately) the change in the relative price of housing, or what economists call the **real price** of housing. It is obtained by subtracting the rate of change in the general price index from the rate of change in the variable of interest.

> **Consumer price index**: the average price level for consumer goods and services.
>
> **Inflation (deflation) rate**: the annual percentage increase (decrease) in the level of consumer prices.
>
> **Real price**: the actual price adjusted by the general (consumer) price level in the economy.

Index numbers

Statistics Canada and other statistical agencies frequently present data in **index number** form. An index number provides an easy way to read the data. For example, suppose we wanted to compute the percentage change in the price of housing between 2001 and 2007. We could do this by entering the two data points in a spreadsheet or calculator and do the computation. But suppose the prices were entered in another form. In particular, by dividing each price value by the first year value and multiplying the result by 100 we obtain a series of prices that are all relative to the initial year – which we call the base year. The resulting series in column 6 of Table 2.1 is an index of house price values. Each entry is the corresponding value in column 2, divided by the first entry in column 2. For example, the value 124.0 in row 4 is obtained as $(434,000/350,000)*100 = 124.0$. The key characteristics of indexes are that they are *not dependent upon the units of measurement of the data in question*, and they are interpretable easily with reference to a given base value. To illustrate, suppose we wish to know how prices behaved between 2001 and 2007. The index number column immediately tells us that prices increased by 80%, because relative to 2001, the 2007 value is 80% higher.

> **Index number**: value for a variable, or an average of a set of variables, expressed relative to a given base value.

Furthermore, index numbers enable us to make *comparisons with the price patterns for other goods* much more easily. If we had constructed a price index for automobiles, which also had a base value of 100 in 2001, we could make immediate comparisons without having to compare one set of numbers defined in thousands of dollars with another defined in hundreds of thousands of dollars. In short, index numbers simplify the interpretation of data.

2.2 Data, theory and economic models

Let us now investigate the interplay between economic theories on the one hand and data on the other. We will develop two examples. The first will be based upon the data on house prices, the second upon a new data set.

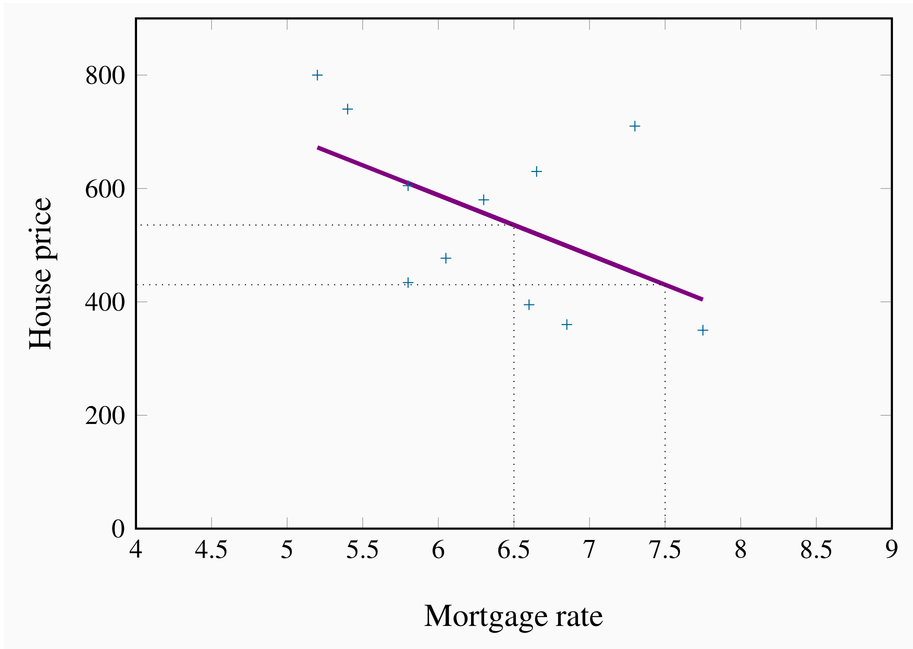
House prices – theory

Remember from Chapter 1 that a theory is a logical argument regarding economic relationships. A theory of house prices would propose that the price of housing depends upon a number of elements in the economy. In particular, if borrowing costs are low then buyers are able to afford the interest costs on larger borrowings. This in turn might mean they are willing to pay higher prices. Conversely, if borrowing rates are higher. Consequently, the borrowing rate, or mortgage rate, is a variable for an economic model of house prices. A second variable might be available space for development: If space in a given metropolitan area is tight then the land value will reflect this, and consequently the higher land price should be reflected in higher house prices. A third variable would be the business climate: If there is a high volume of high-value business transacted in a given area then buildings will be more in demand, and that in turn should be reflected in higher prices. For example, both business and residential properties are more highly priced in San Francisco and New York than in Moncton, New Brunswick. A fourth variable might be environmental attractiveness: Vancouver may be more enticing than other towns in Canada. A fifth variable might be the climate.

House prices – evidence

These and other variables could form the basis of a *theory* of house prices. A *model* of house prices, as explained in Chapter 1, focuses upon what we would consider to be the most important subset of these variables. In the limit, we could have an extremely simple model that specified a dependence between the price of housing and the mortgage rate alone. To test such a simple model we need data on house prices and mortgage rates. The final column of Table 2.1 contains data on the 5-year fixed-rate mortgage for the period in question. Since our simple model proposes that prices depend (primarily) upon mortgage rates, in Figure 2.3 we plot the house price series on the vertical axis, and the mortgage rate on the horizontal axis, for each year from 2001 to 2011. As before, each point (shown as a '+') represents a pair of price and mortgage rate values.

Figure 2.3: Price of housing



The resulting plot (called a scatter diagram) suggests that there is a negative relationship between these two variables. That is, higher prices are correlated with lower mortgage rates. Such a correlation is consistent with our theory of house prices, and so we might conclude that changes in mortgage rates *cause* changes in house prices. Or at least the data suggest that we should not reject the idea that such causation is in the data.

House prices – inference

To summarize the relationship between these variables, the pattern suggests that a straight line through the scatter plot would provide a reasonably good description of the relationship between these variables. Obviously it is important to define the most appropriate line – one that 'fits' the data well.[1] The line we have drawn through the data points is informative, because it relates the two variables in a *quantitative manner*. It is called a **regression line**. It predicts that, on average, if the mortgage rate increases, the price of housing will respond in the downward direction. This particular line states that a one point change in the mortgage rate will move prices in the opposing direction by $105,000. This is easily verified by considering the dollar value corresponding to say a mortgage value of 6.5, and then the value corresponding to a mortgage value of 7.5. Projecting vertically to the regression line from each of these points on the horizontal axis, and from there across to the vertical axis will produce a change in price of $105,000.

[1]This task is the job of econometricians, who practice econometrics. Econometrics is the science of examining and quantifying relationships between economic variables. It attempts to determine the separate influences of each variable, in an environment where many things move simultaneously. Computer algorithms that do this are plentiful. Computers can also work in many dimensions in order to capture the influences of *several variables simultaneously* if the model requires that.

Note that the line is not at all a 'perfect' fit. For example, the mortgage rate declined between 2008 and 2009, but the price declined also – contrary to our theory. The model is not a perfect predictor; it states that *on average* a change in the magnitude of the x-axis variable leads to a change of a specific amount in the magnitude of the y-axis variable.

In this instance the slope of the line is given by -105,000/1, which is the vertical distance divided by the corresponding horizontal distance. Since the line is straight, this slope is unchanging.

> **Regression line**: representation of the average relationship between two variables in a scatter diagram.

Road fatalities – theory, evidence and inference

Table 2.3 contains data on annual road fatalities per 100,000 drivers for various age groups. In the background, we have a *theory*, proposing that driver fatalities depend upon the age of the driver, the quality of roads and signage, speed limits, the age of the automobile stock and perhaps some other variables. Our model focuses upon a subset of these variables, and in order to present the example in graphical terms we specify fatalities as being dependent upon a single variable – age of driver.

Table 2.3: Non-linearity: Driver fatality rates Canada, 2009

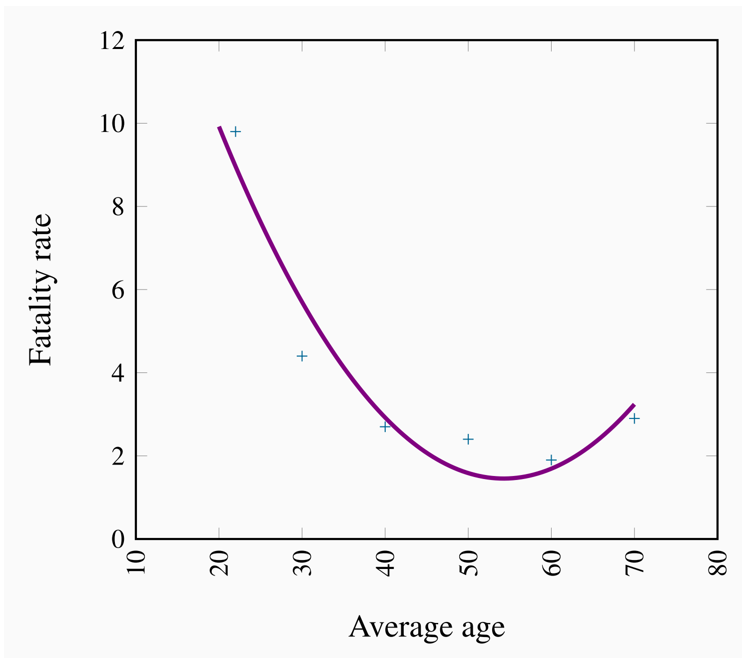
Age of driver	Fatality rate per 100,000 drivers
20-24	9.8
25-34	4.4
35-44	2.7
45-54	2.4
55-64	1.9
65+	2.9

Source: Transport Canada, Canadian motor vehicle traffic collision statistics, 2009.

The scatter diagram is presented in Figure 2.4. Two aspects of this plot stand out. First, there is an exceedingly steep decline in the fatality rate when we go from the youngest age group to the next two age groups. The decline in fatalities between the youngest and second youngest groups is 5.4 points, and between the second and third age groups is 1.7 points. The decline between the third and fourth groups is minimal - just 0.3 points. Hence, behaviour is not the same throughout the age distribution. Second, we notice that fatalities increase for the oldest age group, perhaps indicating that the oldest drivers are not as good as middle-aged drivers.

These two features suggest that the relationship between fatalities and age differs across the age spectrum. Accordingly, a straight line would not be an accurate way of representing the behaviours in these data. A straight line through the plot implies that a given change in age should have a similar impact on fatalities, no matter the age group. Accordingly we have an example of a *non-linear relationship*. Such a non-linear relationship might be represented by the curve going through the plot. Clearly the slope of this line varies as we move from one age category to another.

Figure 2.4: Non-linearity: Driver fatality rates Canada, 2009



Fatality rates vary non-linearly with age: At first they decline, then increase again, relative to the youngest age group.

2.3 Ethics, efficiency and beliefs

Positive economics studies objective or scientific explanations of how the economy functions. Its aim is to understand and generate predictions about how the economy may respond to changes and policy initiatives. In this effort economists strive to act as detached scientists, regardless of political sympathies or ethical code. Personal judgments and preferences are (ideally) kept apart. In this particular sense, economics is similar to the natural sciences such as physics or biology. To date in this chapter we have been exploring economics primarily from a positive standpoint.

In contrast, **normative economics** offers recommendations based partly on value judgments. While economists of different political persuasions can agree that raising the income tax rate would lead to some reduction in the number of hours worked, they may yet differ in their views on the advisability of such a rise. One economist may believe that the additional revenue that may come in to government coffers is not worth the disincentives to work; another may think that, if such monies can be redistributed to benefit the needy, or provide valuable infrastructure, the negative impact on the workers paying the income tax is worth it.

> **Positive economics** studies objective or scientific explanations of how the economy functions.
>
> **Normative economics** offers recommendations that incorporate value judgments.

Scientific research can frequently resolve differences that arise in positive economics—not so in normative economics. For example, if we claim that "the elderly have high medical bills, and the government should cover all of the bills", we are making both a positive and a normative statement. The first part is positive, and its truth is easily established. The latter part is normative, and individuals of different beliefs may reasonably differ. Some people may believe that the money would be better spent on the environment and have the aged cover at least part of their own medical costs. Positive economics does not attempt to show that one of these views is correct and the other false. The views are based on value judgments, and are motivated by a concern for **equity**. Equity is a vital guiding principle in the formation of policy and is frequently, though not always, seen as being in competition with the drive for economic growth. Equity is driven primarily by normative considerations. Few economists would disagree with the assertion that a government should implement policies that improve the lot of the poor—but to what degree?

> **Economic equity** is concerned with the distribution of well-being among members of the economy.

Application Box 2.1: Wealth Tax

US Senator Elizabeth Warren, in seeking her (Democratic) Party's nomination as candidate for the Presidency in 2020, proposed that individuals with high wealth should pay a wealth tax. Her proposal was to levy a tax of 2% on individual wealth holdings above $50 million and a 6% tax on wealth above one billion dollars. This is clearly a *normative* approach to the issue of wealth concentration; it represented her ethical solution to what she perceived as socially unjust inequality.

In contrast, others in her party (Professor Larry Summers of Harvard for example) argued that the impact of such a tax would be to incentivize wealthy individuals to reclassify their wealth, or to give it to family members, or to offshore it, in order to avoid such a tax. If individuals behaved in this way the tax take would be far less than envisaged by Senator Warren. Such an analysis by Professor Summers is *positive* in nature; it attempts to define what might happen in response to the normative policy of Senator Warren. If he took the further step of saying that wealth should not be taxed, then he would be venturing into normative territory. Henry Aaron of the Brookings Institution argued that a more progressive inheritance tax than currently exists would be easier to implement, and would be more effective in both generating tax revenue and equalizing wealth holdings. Since he also advocated implementing such a proposal, as an alternative to Senator Warren's proposals, he was being both *normative and positive*.

Most economists hold normative views, sometimes very strongly. They frequently see themselves, not just as cold hearted scientists, but as champions for their (normative) cause in addition. Conservative economists see a smaller role for government than left-leaning economists.

Many economists see a conflict between equity and the efficiency considerations that we developed in Chapter 1. For example, high taxes may provide disincentives to work in the marketplace and therefore reduce the efficiency of the economy: Plumbers and gardeners may decide to do their own gardening and their own plumbing because, by staying out of the marketplace where monetary transactions are taxed, they can avoid the taxes. And avoiding the taxes may turn out to be as valuable as the efficiency gains they forgo.

In other areas the equity-efficiency trade-off is not so obvious: If taxes (that may have disincentive effects) are used to educate individuals who otherwise would not develop the skills that follow education, then economic growth may be higher as a result of the intervention.

Revisiting the definition of economics – core beliefs

This is an appropriate point at which to return to the definition of economics in Chapter 1 that we borrowed from Nobel Laureate Christopher Sims: Economics is a set of ideas and methods for the betterment of society.

If economics is concerned about the betterment of society, clearly there are ethical as well as efficiency considerations at play. And given the philosophical differences among scientists (including economists), can we define an approach to economics that is shared by the economics profession at large? Most economists would answer that the profession shares a set of beliefs, and that differences refer to the extent to which one consideration may collide with another.

- First of all we believe that *markets are critical* because they facilitate exchange and therefore encourage efficiency. Specialization and trade creates benefits for the trading parties. For example, Canada has not the appropriate climate for growing coffee beans, and Colombia has not the terrain for wheat. If Canada had to be self-sufficient, we might have to grow coffee beans in green-houses—a costly proposition. But with trade we can specialize, and then exchange some of our wheat for Colombian coffee. Similar benefits arise for the Colombians.

  A frequent complaint against trade is that its modern-day form (globalization) does not benefit the poor. For example, workers in the Philippines may earn only a few dollars per day manufacturing clothing for Western markets. From this perspective, most of the gains from trade go to the Western consumers and capitalists, come at the expense of jobs to western workers, and provide Asian workers with meagre rewards.

- A corollary of the centrality of markets is *that incentives matter*. If the price of business class seats on your favourite airline is reduced, you may consider upgrading. Economists believe that the price mechanism influences behaviour, and therefore favour the use of price incentives in the marketplace and public policy more generally. Environmental economists, for example, advocate the use of pollution permits that can be traded at a price between users, or carbon taxes on the emission of greenhouse gases. We will develop such ideas in *Principles of Microeconomics* Chapter 5 more fully.

- In saying that economists believe in incentives, we are not proposing that human beings are purely mercenary. People have many motivations: Self-interest, a sense of public duty, kindness, etc. Acting out of a sense of self-interest does not imply that people are morally empty or have no altruistic sense.
- Economists believe universally in the *importance of the rule of law*, no matter where they sit on the political spectrum. Legal institutions that govern contracts are critical to the functioning of an economy. If goods and services are to be supplied in a market economy, the suppliers must be guaranteed that they will be remunerated. And this requires a developed legal structure with penalties imposed on individuals or groups who violate contracts. Markets alone will not function efficiently.

  Modern development economics sees the implementation of the rule of law as perhaps the central challenge facing poorer economies. There is a strong correlation between economic growth and national wealth on the one hand, and an effective judicial and policing system on the other. The consequence on the world stage is that numerous 'economic' development projects now focus upon training jurists, police officers and bureaucrats in the rule of law!
- Finally, economists believe in the centrality of government. Governments can solve a number of problems that arise in market economies that cannot be addressed by the private market place. For example, governments can best address the potential abuses of monopoly power. Monopoly power, as we shall see in *Microeconomics* Chapter 10, not only has equity impacts it may also reduce economic efficiency. Governments are also best positioned to deal with environmental or other types of externalities – the impact of economic activity on sectors of the economy that are not directly involved in the activity under consideration.

In summary, governments have a variety of roles to play in the economy. These roles involve making the economy more equitable and more efficient by using their many powers.

KEY TERMS

Variables: measures that can take on different sizes.

Data: recorded values of variables.

Time series data: a set of measurements made sequentially at different points in time.

High (low) frequency data series have short (long) intervals between observations.

Cross-section data: values for different variables recorded at a point in time.

Repeated cross-section data: cross-section data recorded at regular or irregular intervals.

Longitudinal data follow the same units of observation through time.

Percentage change$=$ (change in values)$/$original value $\times$ 100.

Consumer price index: the average price level for consumer goods and services.

Inflation (deflation) rate: the annual percentage increase (decrease) in the level of consumer prices.

Real price: the actual price adjusted by the general (consumer) price level in the economy.

Index number: value for a variable, or an average of a set of variables, expressed relative to a given base value.

Regression line: representation of the average relationship between two variables in a scatter diagram.

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

Normative economics offers recommendations that incorporate value judgments.

Economic equity is concerned with the distribution of well-being among members of the economy.

Exercises for Chapter 2

Exercise 2.1 An examination of a country's recent international trade flows yields the data in the table below.

Year	National Income ($b)	Imports ($b)
2011	1,500	550
2012	1,575	573
2013	1,701	610
2014	1,531	560
2015	1,638	591

(a) Based on an examination of these data do you think the national income and imports are not related, positively related, or negatively related?

(b) Plot each pair of observations in a two-dimensional line diagram to illustrate your view of the import/income relationship. Measure income on the horizontal axis and imports on the vertical axis. This can be done using graph paper or a spreadsheet-cum-graphics software.

Exercise 2.2 The average price of a medium coffee at *Wakeup Coffee Shop* in each of the past ten years is given in the table below.

2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
$1.05	$1.10	$1.14	$1.20	$1.25	$1.25	$1.33	$1.35	$1.45	$1.49

(a) Construct an annual 'coffee price index' for this time period using 2005 as the base year. [*Hint*: follow the procedure detailed in the chapter – divide each yearly price by the base year price.]

(b) Based on your price index, what was the percentage change in the price of a medium coffee from 2005 to 2012?

(c) Based on your index, what was the average annual percentage change in the price of coffee from 2005 to 2010?

(d) Assuming the inflation rate in this economy was 2% every year, what was the real change in the price of coffee between 2007 and 2008; and between 2009 and 2010?

Exercise 2.3 The following table shows hypothetical consumption spending by households and income of households in billions of dollars.

Year	Income	Consumption
2006	476	434
2007	482	447
2008	495	454
2009	505	471
2010	525	489
2011	539	509
2012	550	530
2013	567	548

(a) Plot the scatter diagram with consumption on the vertical axis and income on the horizontal axis.

(b) Fit a line through these points.

(c) Does the line indicate that these two variables are related to each other?

(d) How would you describe the *causal relationship* between income and consumption?

Exercise 2.4 Using the data from Exercise 2.3, compute the percentage change in consumption and the percentage change in income for each pair of adjoining years between 2006 and 2013.

Exercise 2.5 You are told that the relationship between two variables, X and Y, has the form $Y = 10 + 2X$. By trying different values for X you can obtain the corresponding predicted value for Y (e.g., if $X = 3$, then $Y = 10 + 2 \times 3 = 16$). For values of X between 0 and 12, compute the matching value of Y and plot the scatter diagram.

Exercise 2.6 For the data below, plot a scatter diagram with variable Y on the vertical axis and variable X on the horizontal axis.

Y	40	33	29	56	81	19	20
X	5	7	9	3	1	11	10

(a) Is the relationship between the variables positive or negative?

(b) Do you think that a linear or non-linear line better describes the relationship?

Chapter 3

The classical marketplace – demand and supply

In this chapter we will explore:

3.1 The marketplace – trading

The marketplace in today's economy has evolved from earlier times. It no longer has a unique form – one where buyers and sellers physically come together for the purpose of exchange. Indeed, supermarkets usually require individuals to be physically present to make their purchases. But when purchasing an airline ticket, individuals simply go online and interact with perhaps a number of different airlines (suppliers) simultaneously. Or again, individuals may simply give an instruction to their stock broker, who will execute a purchase on their behalf – the broker performs the role of a middleman, who may additionally give advice to the purchaser. Or a marketing agency may decide to subcontract work to a translator or graphic artist who resides in Mumbai. The advent of the coronavirus has shifted grocery purchases from in-store presence to home delivery for many buyers. In pure auctions (where a single work of art or a single residence is offered for sale) buyers compete one against the other for the single item supplied. Accommodations in private homes are supplied to potential visitors (buyers) through *Airbnb*. Taxi rides are mediated through *Lyft* or *Uber*. These institutions are all different types of markets; they serve the purpose of facilitating exchange and trade.

Not all goods and services in the modern economy are obtained through the marketplace. Schooling and health care are allocated in Canada primarily by government decree. In some instances the market plays a supporting role: Universities and colleges may levy fees, and most individuals must pay, at least in part, for their pharmaceuticals. In contrast, broadcasting services may carry a price of zero; Buzzfeed or other news and social media come free of payment. Furthermore, some markets have no price, yet they find a way of facilitating an exchange. For example, graduating medical students need to be matched with hospitals for their residencies. Matching mechanisms are a form of market in that they bring together suppliers and demanders. We explore their operation in Chapter 11.

The importance of the marketplace springs from its role as an allocating mechanism. Elevated prices effectively send a signal to suppliers that the buyers in the market place a high value on the product being traded; conversely when prices are low. Accordingly, suppliers may decide to cease supplying markets where prices do not remunerate them sufficiently, and redirect their energies and the productive resources under their control to other markets – markets where the product being traded is more highly valued, and where the buyer is willing to pay more.

Whatever their form, the marketplace is central to the economy we live in. Not only does it facilitate trade, it also provides a means of earning a livelihood. Suppliers must hire resources – human and non-human – in order to bring their supplies to market and these resources must be paid a return: income is generated.

In this chapter we will examine the process of price formation – how the prices that we observe in the marketplace come to be what they are. We will illustrate that the price for a good is inevitably linked to the quantity of a good; price and quantity are different sides of the same coin and cannot generally be analyzed separately. To understand this process more fully, we need to *model* a typical market. The essentials are demand and supply.

3.2 The market's building blocks

In economics we use the terminology that describes trade in a particular manner. Non-economists frequently describe microeconomics by saying "it's all about supply and demand". While this is largely true we need to define exactly what we mean by these two central words. **Demand** is the quantity of a good or service that buyers wish to purchase at each conceivable price, with all other influences on demand remaining unchanged. It reflects a multitude of values, not a single value. It is not a single or unique quantity such as two cell phones, but rather a full description of the quantity of a good or service that buyers would purchase at various prices.

> **Demand** is the quantity of a good or service that buyers wish to purchase at each possible price, with all other influences on demand remaining unchanged.

As a hypothetical example, the first column of Table 3.1 shows the price of natural gas per cubic foot. The second column shows the quantity that would be purchased in a given time period at each price. It is therefore a schedule of quantities demanded at various prices. For example, at a price \$6 per unit, buyers would like to purchase 4 units, whereas at the lower price of \$3 buyers would like to purchase 7 units. Note also that this is a homogeneous good. A cubit foot of natural gas is considered to be the same product no matter which supplier brings it to the market. In contrast, accommodations supplied through *Airbnb* are heterogeneous – they vary in size and quality.

Table 3.1: Demand and supply for natural gas

Price ($)	Demand (thousands of cu feet)	Supply (thousands of cu feet)	Excess
10	0	18	
9	1	16	
8	2	14	**Excess Supply**
7	3	12	
6	4	10	
5	5	8	
4	6	6	**Equilibrium**
3	7	4	
2	8	2	**Excess Demand**
1	9	0	
0	10	0	

Supply is interpreted in a similar manner. It is not a single value; we say that **supply** is the quantity of a good or service that sellers are willing to sell at each possible price, with all other influences on supply remaining unchanged. Such a supply schedule is defined in the third column of the table. It is assumed that no supplier can make a profit (on account of their costs) unless the price is at least $2 per unit, and therefore a zero quantity is supplied below that price. The higher price is more profitable, and therefore induces a greater quantity supplied, perhaps by attracting more suppliers. This is reflected in the data. For example, at a price of $3 suppliers are willing to supply 4 units, whereas with a price of $7 they are willing to supply 12 units. There is thus a positive relationship between price and quantity for the supplier – a higher price induces a greater quantity; whereas on the demand side of the market a higher price induces a lower quantity demanded – a negative relationship.

> **Supply** is the quantity of a good or service that sellers are willing to sell at each possible price, with all other influences on supply remaining unchanged.

We can now identify a key difference in terminology – between the words demand and quantity demanded, and between supply and quantity supplied. While the words demand and supply refer to the complete schedules of demand and supply, the terms **quantity demanded** and **quantity supplied** each define a single value of demand or supply at a particular price.

> **Quantity demanded** defines the amount purchased at a particular price.
>
> **Quantity supplied** refers to the amount supplied at a particular price.

Thus while the non-economist may say that when some fans did not get tickets to the Stanley Cup it was a case of demand exceeding supply, as economists we say that the quantity demanded exceeded the quantity supplied *at the going price of tickets*. In this instance, had every ticket been offered at a sufficiently high price, the market could have generated an excess supply rather than an excess demand. A higher ticket price would reduce the *quantity demanded*; yet would not change *demand*, because demand refers to the whole schedule of possible quantities demanded at different prices.

Other things equal – ceteris paribus

The demand and supply schedules rest on the assumption that all other influences on supply and demand remain the same as we move up and down the possible price values. The expression *other things being equal*, or its Latin counterpart *ceteris paribus*, describes this constancy of other influences. For example, we assume on the demand side that the prices of other goods remain constant, and that tastes and incomes are unchanging. On the supply side we assume, for example, that there is no technological change in production methods. If any of these elements change then the market supply or demand schedules will reflect such changes. For example, if coal or oil prices increase (decline) then some buyers may switch to (away from) gas or solar power. This will be reflected in the data: At any given price more (or less) will be demanded. We will illustrate this in graphic form presently.

Market equilibrium

Let us now bring the demand and supply schedules together in an attempt to analyze what the marketplace will produce – will a single price emerge that will equate supply and demand? We will keep other things constant for the moment, and explore what materializes at different prices. At low prices, the data in Table 3.1 indicate that the quantity demanded exceeds the quantity supplied – for example, verify what happens when the price is \$3 per unit. The opposite occurs when the price is high – what would happen if the price were \$8? Evidently, there exists an intermediate price, where the quantity demanded equals the quantity supplied. At this point we say that the market is in equilibrium. The **equilibrium price** equates demand and supply – it clears the market.

> The **equilibrium price** equilibrates the market. It is the price at which quantity demanded equals the quantity supplied.

In Table 3.1 the equilibrium price is \$4, and the equilibrium quantity is 6 thousand cubic feet of gas (we use the notation ‘k’ to denote thousands). At higher prices there is an **excess supply**—suppliers wish to sell more than buyers wish to buy. Conversely, at lower prices there is an **excess demand**. Only at the equilibrium price is the quantity supplied equal to the quantity demanded.

> **Excess supply** exists when the quantity supplied exceeds the quantity demanded at the going price.
>
> **Excess demand** exists when the quantity demanded exceeds the quantity supplied at the going price.

Does the market automatically reach equilibrium? To answer this question, suppose initially that the sellers choose a price of \$10. Here suppliers would like to supply 18k cubic feet, but there are no buyers—a situation of extreme excess supply. At the price of \$7 the excess supply is reduced to 9k, because both the quantity demanded is now higher at 3k units, and the quantity supplied is lower at 12k. But excess supply means that there are suppliers willing to supply at a lower price, and this willingness exerts continual downward pressure on any price above the price that equates demand and supply.

At prices below the equilibrium there is, conversely, an excess demand. In this situation, suppliers could force the price upward, knowing that buyers will continue to buy at a price at which the suppliers are willing to sell. Such upward pressure would continue until the excess demand is eliminated.

In general then, above the equilibrium price excess supply exerts downward pressure on price, and below the equilibrium excess demand exerts upward pressure on price. This process implies that the buyers and sellers have information on the various elements that make up the marketplace.

We will explore later in this chapter some specific circumstances in which trading could take place at prices above or below the equilibrium price. In such situations the quantity actually traded always corresponds to the short side of the market: this means that at high prices the quantity demanded is less than the quantity supplied, and it is the quantity demanded that is traded because buyers will not buy the amount suppliers would like to supply. Correspondingly, at low prices the quantity demanded exceeds quantity supplied, and it is the amount that suppliers are willing to sell that is traded. In sum, when trading takes place at prices other than the equilibrium price it is always the lesser of the quantity demanded or supplied that is traded. Hence we say that at non-equilibrium prices the **short side** dominates. We will return to this in a series of examples later in this chapter.

> The **short side of the market** determines outcomes at prices other than the equilibrium.

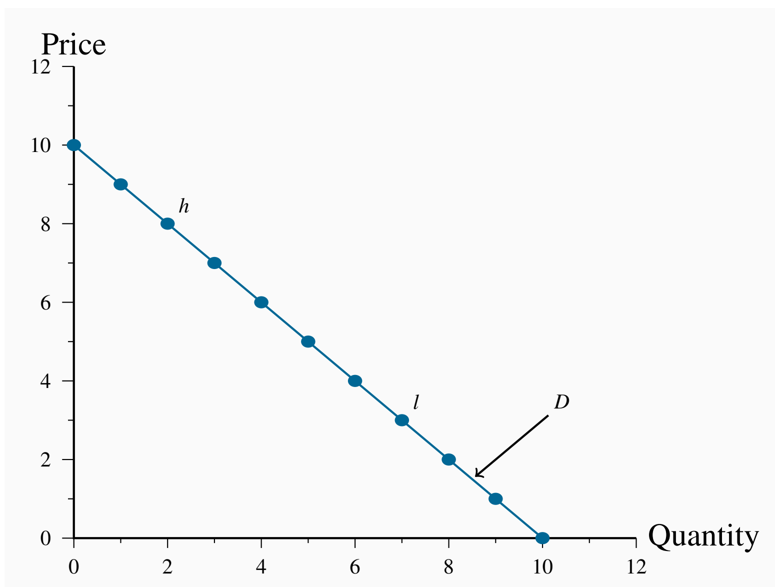
Supply and the nature of costs

Before progressing to a graphical analysis, we should add a word about costs. The supply schedules are based primarily on the cost of producing the product in question, and we frequently assume that all of the costs associated with supply are incorporated in the supply schedules. In Chapter 6 we will explore cases where costs additional to those incurred by producers may be relevant. For example, coal burning power plants emit pollutants into the atmosphere; but the individual supplier may not take account of these pollutants, which are costs to society at large, in deciding how much to supply at different prices. Stated another way, the private costs of production would not reflect the total, or full social costs of production. Conversely, if some individuals immunize themselves against a rampant virus, other individuals gain from that action because they become less likely to contract the virus - the social value thus exceeds the private value. For the moment the assumption is that no such additional costs are associated with the markets we analyze.

3.3 Demand and supply curves

The **demand curve** is a graphical expression of the relationship between price and quantity demanded, holding other things constant. Figure 3.1 measures price on the vertical axis and quantity on the horizontal axis. The curve D represents the data from the first two columns of Table 3.1. Each combination of price and quantity demanded lies on the curve. In this case the curve is *linear*—it is a straight line. The demand curve slopes downward (technically we say that its slope is negative), reflecting the fact that buyers wish to purchase more when the price is less.

Figure 3.1: Measuring price & quantity



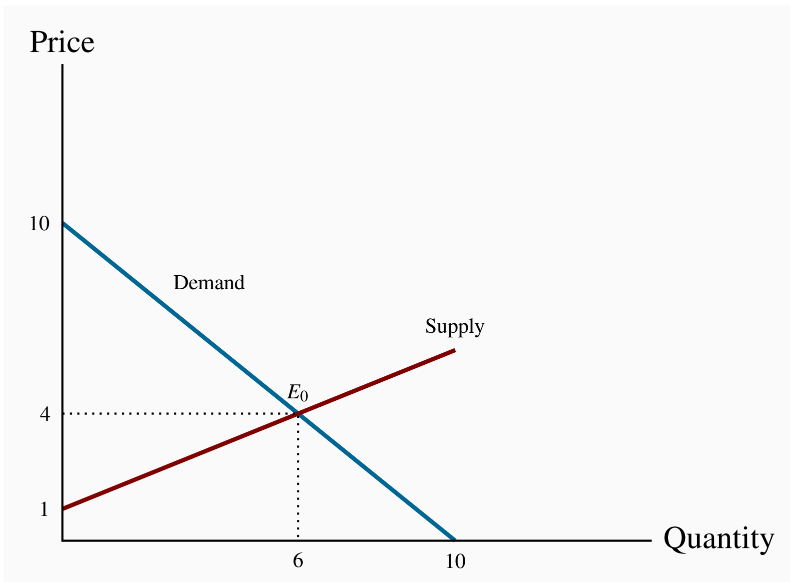
To derive this demand curve we take each price-quantity combination from the demand schedule in Table 3.1 and insert a point that corresponds to those combinations. For example, point h defines the combination $\{P = \$8, Q_d = 2\}$, the point l denotes the combination $\{P = \$3, Q_d = 7\}$. If we join all such points we obtain the demand curve in Figure 3.2.The same process yields the supply curve in Figure 3.2. In this example the supply and the demand curves are each linear. There is no reason why this linear property characterizes demand and supply curves in the real world; they are frequently found to have curvature. But straight lines are easier to work with, so we continue with them for the moment.

> The **demand curve** is a graphical expression of the relationship between price and quantity demanded, with other influences remaining unchanged.

The **supply curve** is a graphical representation of the relationship between price and quantity supplied, holding other things constant. The supply curve *S* in Figure 3.2 is based on the data from columns 1 and 3 in Table 3.1. It has a positive slope indicating that suppliers wish to supply more at higher prices.

> The **supply curve** is a graphical expression of the relationship between price and quantity supplied, with other influences remaining unchanged.

Figure 3.2: Supply, demand, equilibrium



The demand and supply curves intersect at point E_0, corresponding to a price of \$4 which, as illustrated above, is the equilibrium price for this market. At any price below this the horizontal distance between the supply and demand curves represents excess demand, because demand exceeds supply. Conversely, at any price above \$4 there is an excess supply that is again measured by the horizontal distance between the two curves. Market forces tend to eliminate excess demand and excess supply as we explained above. In the final section of the chapter we illustrate how the supply and demand curves can be 'solved' for the equilibrium price and quantity.

3.4 Non-price influences on demand

We have emphasized several times the importance of the *ceteris paribus* assumption when exploring the impact of different prices on the quantity demanded: We assume all other influences on the purchase decision are unchanged (at least momentarily). These other influences fall into several broad categories: The prices of related goods; the incomes of buyers; buyer tastes; and expectations about the future. Before proceeding, note that we are dealing with *market* demand rather than demand by one *individual* (the precise relationship between the two is developed later in this chapter).

The prices of related goods – oil and gas, Kindle and paperbacks

We expect that the price of other forms of energy would impact the price of natural gas. For example, if hydro-electricity, oil or solar becomes less expensive we would expect some buyers to switch to these other products. Alternatively, if gas-burning furnaces experience a technological breakthrough that makes them more efficient and cheaper we would expect some users of other fuels to move to gas. Among these examples, oil and electricity are substitute fuels for gas; in contrast a more fuel-efficient new gas furnace complements the use of gas. We use these terms, **substitutes** and **complements**, to describe products that influence the demand for the primary good.

> **Substitute goods**: when a price reduction (rise) for a related product reduces (increases) the demand for a primary product, it is a substitute for the primary product.
>
> **Complementary goods**: when a price reduction (rise) for a related product increases (reduces) the demand for a primary product, it is a complement for the primary product.

Clearly electricity is a substitute for gas in the power market, whereas a gas furnace is a complement for gas as a fuel. The words substitutes and complements immediately suggest the nature of the relationships. Every product has complements and substitutes. As another example: Electronic readers and tablets are substitutes for paper-form books; a rise in the price of paper books should increase the demand for electronic readers at any given price for electronic readers. In graphical terms, the demand curve *shifts* in response to changes in the prices of other goods – an increase in the price of paper-form books shifts the demand for electronic readers outward, because more electronic readers will be demanded at any price.

Buyer incomes – which goods to buy

The demand for most goods increases in response to income growth. Given this, the demand curve for gas will shift outward if household incomes in the economy increase. Household incomes may increase either because there are more households in the economy or because the incomes of the existing households grow.

Most goods are demanded in greater quantity in response to higher incomes at any given price. But there are exceptions. For example, public transit demand may decline at any price when household incomes rise, because some individuals move to cars. Or the demand for laundromats may decline in response to higher incomes, as households purchase more of their own consumer durables – washers and driers. We use the term **inferior good** to define these cases: An inferior good is one whose demand declines in response to increasing incomes, whereas a **normal good** experiences an increase in demand in response to rising incomes.

> An **inferior good** is one whose demand falls in response to higher incomes.
>
> A **normal good** is one whose demand increases in response to higher incomes.

There is a further sense in which consumer incomes influence demand, and this relates to how the incomes are *distributed* in the economy. In the discussion above we stated that higher total incomes shift demand curves outwards when goods are normal. But think of the difference in the demand for electronic readers between Portugal and Saudi Arabia. These economies have roughly the same average per-person income, but incomes are distributed more unequally in Saudi Arabia. It does not have a large middle class that can afford electronic readers or *iPads*, despite the huge wealth held by the elite. In contrast, Portugal has a relatively larger middle class that can afford such goods. Consequently, the *distribution of income* can be an important determinant of the demand for many commodities and services.

Tastes and networks – hemlines, lapels and homogeneity

While demand functions are drawn on the assumption that tastes are constant, in an evolving world they are not. We are all subject to peer pressure, the fashion industry, marketing, and a desire to maintain our image. If the fashion industry dictates that lapels on men's suits or long skirts are *de rigueur* for the coming season, some fashion-conscious individuals will discard a large segment of their wardrobe, even though the clothes may be in perfectly good condition: Their demand is influenced by the dictates of current fashion.

Correspondingly, the items that other individuals buy or use frequently determine our own purchases. Businesses frequently decide that all of their employees will have the same type of computer and software on account of *network economies*: It is easier to communicate if equipment is compatible, and it is less costly to maintain infrastructure where the variety is less.

Expectations – betting on the future

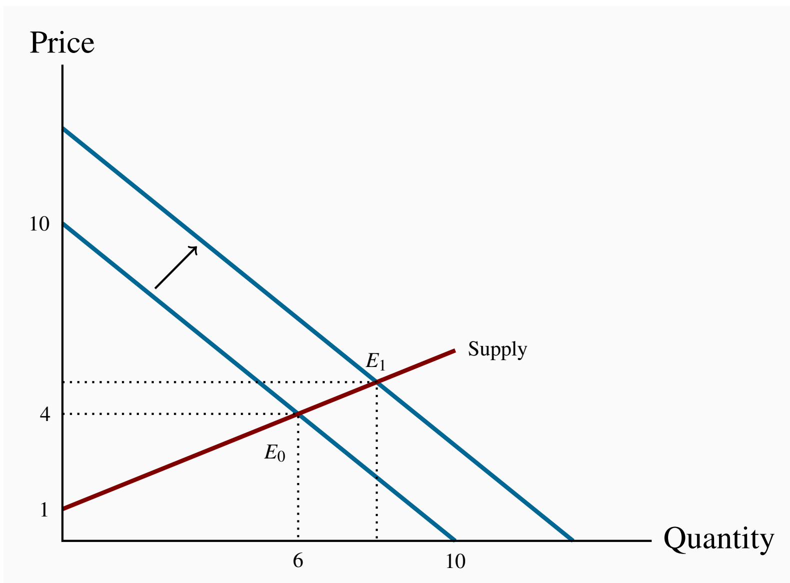
In our natural gas example, if households expected that the price of natural gas was going to stay relatively low for many years – perhaps on account of the discovery of large deposits – then they would be tempted to purchase a gas burning furnace rather than one based upon an alternative fuel. In this example, it is more than the current price that determines choices; *the prices that are expected to prevail in the future* also determine current demand.

Expectations are particularly important in stock markets. When investors anticipate that corporations will earn high rewards in the future they will buy a stock today. If enough people believe this, the price of the stock will be driven upward on the market, even before profitable earnings are registered.

Shifts in demand

The demand curve in Figure 3.2 is drawn for a given level of other prices, incomes, tastes, and expectations. Movements along the demand curve reflect solely the impact of different prices for the good in question, holding other influences constant. But changes in any of these other factors will change the position of the demand curve. Figure 3.3 illustrates a shift in the demand curve. This shift could result from a rise in household incomes that increase the quantity demanded *at every price*. This is illustrated by an outward shift in the demand curve. With supply conditions unchanged, there is a new equilibrium at E_1, indicating a greater quantity of purchases accompanied by a higher price. The new equilibrium reflects a *change in quantity supplied and a change in demand*.

Figure 3.3: Demand shift and new equilibrium



The outward shift in demand leads to a new equilibrium E_1.

We may well ask why so much emphasis in our diagrams and analysis is placed on the relationship between *price* and quantity, rather than on the relationship between quantity and its other determinants. The answer is that we could indeed draw diagrams with quantity on the horizontal axis and a measure of one of these other influences on the vertical axis. But the price mechanism plays a very important role. *Variations in price are what equilibrate the market.* By focusing primarily upon the price, we see the self-correcting mechanism by which the market reacts to excess supply or excess demand.

In addition, this analysis illustrates the method of **comparative statics**—examining the impact of changing one of the other things that are assumed constant in the supply and demand diagrams.

Comparative static analysis compares an initial equilibrium with a new equilibrium, where the difference is due to a change in one of the other things that lie behind the demand curve or the supply curve.

'Comparative' obviously denotes the idea of a comparison, and static means that we are not in a state of motion. Hence we use these words in conjunction to indicate that we compare one outcome with another, without being concerned too much about the transition from an initial equilibrium to a new equilibrium. The transition would be concerned with dynamics rather than statics. In Figure 3.3 we explain the difference between the points E_0 and E_1 by indicating that there has been a change in incomes or in the price of a substitute good. We do not attempt to analyze the details of this move or the exact path from E_0 to E_1.

Application Box 3.1: Corn prices and demand shifts

In the middle of its second mandate, the Bush Administration in the US decided to encourage the production of ethanol – a fuel that is less polluting than gasoline. The target production was 35 billion for 2017 – from a base of 1 billion gallons in 2000. Corn is the principal input in ethanol production. It is also used as animal feed, as a sweetener and as a food for humans. The target was to be met with the help of a subsidy to producers and a tariff on imports of Brazil's sugar-cane based ethanol.

The impact on corn prices was immediate; from a farm-gate price of $2 per bushel in 2005, the price reached the $4 range two years later. In 2012 the price rose temporarily to $7. While other factors were in play - growing incomes and possibly speculation by commodity investors, ethanol is seen as the main price driver: demand for corn increased and the supply could not be increased to keep up with the demand without an increase in price.

The wider impact of these developments was that the prices of virtually all grains increased in tandem with corn: the prices of sorghum and barley increased because of a switch in land use towards corn on account of its profitability.

While farmers benefited from the price rise, consumers – particularly those in less developed economies – experienced a dramatic increase in their basic living costs. Visit the site of the United Nations' Food and Agricultural Organization for an assessment. Since hitting $7 per bushel in 2012, the price has dropped and averaged $3.50 in 2016.

In terms of supply and demand shifts: the demand side has dominated, particularly in the short run. The ethanol drive, combined with secular growth in the demand for food, means that the demand for grains shifted outward faster than the supply. In the period 2013–2016, supply has increased and the price has moderated.

3.5 Non-price influences on supply

To date we have drawn supply curves with an upward slope. Is this a reasonable representation of supply in view of what is frequently observed in markets? We suggested earlier that the various producers of a particular good or service may have different levels of efficiency. If so, only the more efficient producers can make a profit at a low price, whereas at higher prices more producers or suppliers enter the market – producers who may not be as lean and efficient as those who can survive in a lower-price environment. This view of the world yields a positively-sloping supply curve.

As a second example, consider *Uber* or *Lyft* taxi drivers. Some drivers may be in serious need of income and may be willing to drive for a low hourly rate. For other individuals driving may be a secondary source of income, and such drivers are less likely to want to drive unless the hourly wage is higher. Consequently if these ride sharing services need a large number of drivers at any one time it may be necessary to pay a higher wage – *and charge a higher fare to passengers*, to induce more drivers to take their taxis onto the road. This phenomenon corresponds to a positively-sloped supply curve.

In contrast to these two examples, some suppliers simply choose a unique price and let buyers purchase as much as they want at that price. This is the practice of most retailers. For example, the price of *Samsung's Galaxy* is typically fixed, no matter how many are purchased – and tens of millions are sold at a fixed price when a new model is launched. *Apple* also sets a price, and buyers purchase as many as they desire at that price. This practice corresponds to a horizontal supply curve: The price does not vary and the market equilibrium occurs where the demand curve intersects this supply curve.

In yet other situations supply is fixed. This happens in auctions. Bidders at the auction simply determine the price to be paid. At a real estate auction a given property is put on the market and the price is determined by the bidding process. In this case the supply of a single property is represented by a vertical supply at a quantity of 1 unit.

Regardless of the type of market we encounter, however, it is safe to assume that supply curves rarely slope downward. So, for the moment, we adopt the stance that supply curves are generally upward sloping – somewhere between the extremes of being vertical or horizontal – as we have drawn them to this point.

Next, we examine those other influences that underlie supply curves. Technology, input costs, the prices of competing goods, expectations and the number of suppliers are the most important.

Technology – computers and fracking

A technological advance may involve an idea that allows more output to be produced with the same inputs, or an equal output with fewer inputs. A good example is *just-in-time* technology. Before the modern era, virtually all manufacturers kept large stocks of components in their production facilities, but developments in communications and computers at that time made it possible for manufacturers to link directly with their input suppliers. Nowadays auto assembly plants place their order for, say, seat delivery to their local seat supplier well ahead of assembly time. The seats swing into the assembly area hours or minutes before assembly—just in time. The result is that the assembler reduces her seat inventory (an input) and thereby reduces production cost.

Such a technology-induced cost saving is represented by moving the supply curve downward or outward: The supplier is now able and willing to supply the same quantity at a lower price because of the technological innovation. Or, saying the same thing slightly differently, suppliers will supply more at a given price than before.

A second example relates to the extraction of natural gas. The development of 'fracking' means that companies involved in gas recovery can now do so at a lower cost. Hence they are willing to supply any given quantity at a lower price. A third example concerns aluminum cans. Today they weigh a fraction of what they weighed 20 years ago. This is a technology-based cost saving.

Input costs

Input costs can vary independently of technology. For example, a wage negotiation that grants workers a substantial pay raise will increase the cost of production. This is reflected in a leftward, or upward, supply shift: Any quantity supplied is now priced higher; alternatively, suppliers are willing to supply less at the going price.

Production costs may increase as a result of higher required standards in production. As governments implement new safety or product-stress standards, costs may increase. In this instance the increase in costs is not a 'bad' outcome for the buyer. She may be purchasing a higher quality good as a result.

Competing products – Airbnb versus hotels

If competing products improve in quality or fall in price, a supplier may be forced to follow suit. For example, *Asus* and *Dell* are constantly watching each other's pricing policies. If *Dell* brings out a new generation of computers at a lower price, *Asus* may lower its prices in turn—which is to say that Asus' supply curve will shift downward. Likewise, *Samsung* and *Apple* each responds to the other's pricing and technology behaviours. The arrival of new products in the marketplace also impacts the willingness of suppliers to supply goods at a given price. New intermediaries such as *Airbnb* and *Vacation Rentals by Owner* have shifted the supply curves of hotel rooms downward.

These are some of the many factors that influence the position of the supply curve in a given market.

Application Box 3.2: The price of light

Technological developments have had a staggering impact on many price declines. Professor William Nordhaus of Yale University is an expert on measuring technological change. He has examined the trend in the real price of lighting. Originally, light was provided by whale oil and gas lamps and these sources of lumens (the scientific measure of the amount of light produced) were costly. In his research, Professor Nordhaus pieced together evidence on the actual historic cost of light produced at various times, going all the way back to 1800. He found that light in 1800 cost about 100 times more than in 1900, and light in the year 2000 was a fraction of its cost in 1900. A rough calculation suggests that light was five hundred times more expensive at the start of this 200-year period than at the end, and this was before the arrival of LEDs.

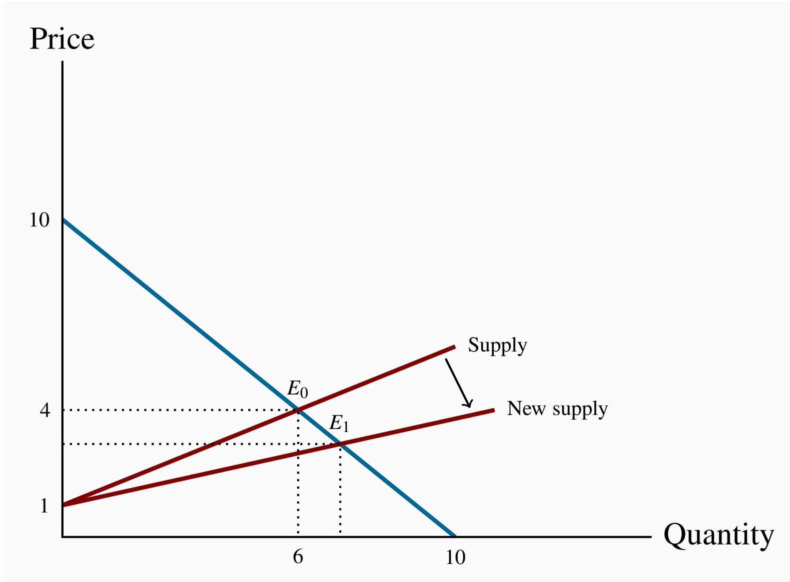
In terms of supply and demand analysis, light has been subject to very substantial downward supply shifts. Despite the long-term growth in demand, the technologically-induced supply changes have been the dominant factor in its price determination.

For further information, visit Professor Nordhaus's website in the Department of Economics at Yale University.

Shifts in supply

Whenever technology changes, or the costs of production change, or the prices of competing products adjust, then one of our *ceteris paribus* assumptions is violated. Such changes are generally reflected by shifting the supply curve. Figure 3.4 illustrates the impact of the arrival of just-in-time technology. The supply curve shifts, reflecting the ability of suppliers to supply the same output at a reduced price. The resulting new equilibrium price is lower, since production costs have fallen. At this reduced price more gas is traded at a lower price.

Figure 3.4: Supply shift and new equilibrium



The supply curve shifts due to lower production costs. A new equilibrium E_1 is attained in the market at a lower price.

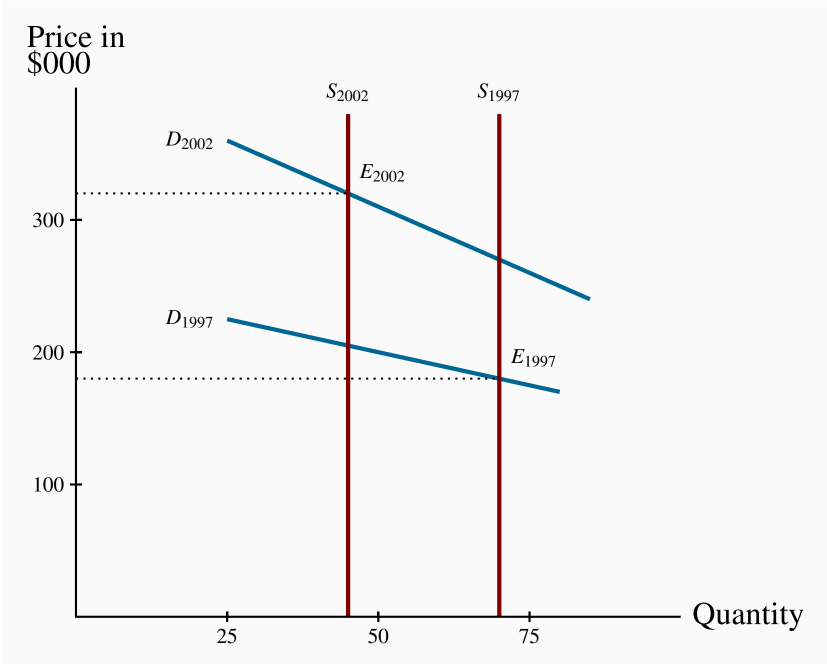
3.6 Simultaneous supply and demand impacts

In the real world, demand and supply frequently shift at the same time. We present such a case in Figure 3.5. It is based upon real estate data describing the housing market in a small Montreal municipality. Vertical curves define the supply side of the market. Such vertical curves mean that a given number of homeowners decide to put their homes on the market, and these suppliers just take whatever price results in the market. In this example, fewer houses were offered for sale in 2002 (less than 50) than in 1997 (more than 70). We are assuming in this market that the houses traded were similar; that is, we are not lumping together mansions with row houses.

During this time period household incomes increased substantially and, also, mortgage rates fell. Both of these developments shifted the demand curve upward/outward: Buyers were willing to pay more for housing in 2002 than in 1997, both because their incomes were on average higher and because they could borrow more cheaply.

The shifts on both sides of the market resulted in a higher average price. And each of these shifts compounded the other: The outward shift in demand would lead to a higher price on its own, and a reduction in supply would do likewise. Hence both forces acted to push up the price in 2002. If, instead, the supply had been greater in 2002 than in 1997 this would have acted to reduce the equilibrium price. And with the demand and supply shifts operating in opposing directions, it is not possible to say in general whether the price would increase or decrease. If the demand shift were strong and the supply shift weak then the demand forces would have dominated and led to a higher price. Conversely, if the supply forces were stronger than the demand forces.

Figure 3.5: A model of the housing market with shifts in demand and supply



The vertical supply denotes a fixed number of houses supplied each year. Demand was stronger in 2002 than in 1997 both on account of higher incomes and lower mortgage rates. Thus the higher price in 2002 is due to both a reduction in supply and an increase in demand.

3.7 Market interventions – governments and interest groups

The freely functioning markets that we have developed certainly do not describe all markets. For example, minimum wages characterize the labour market, most agricultural markets have supply restrictions, apartments are subject to rent controls, and blood is not a freely traded market commodity in Canada. In short, price controls and quotas characterize many markets. **Price controls** are government rules or laws that inhibit the formation of market-determined prices. **Quotas** are physical restrictions on how much output can be brought to the market.

> **Price controls** are government rules or laws that inhibit the formation of market-determined prices.
>
> **Quotas** are physical restrictions on output.

Price controls come in the form of either *floors* or *ceilings*. Price floors are frequently accompanied by *marketing boards*.

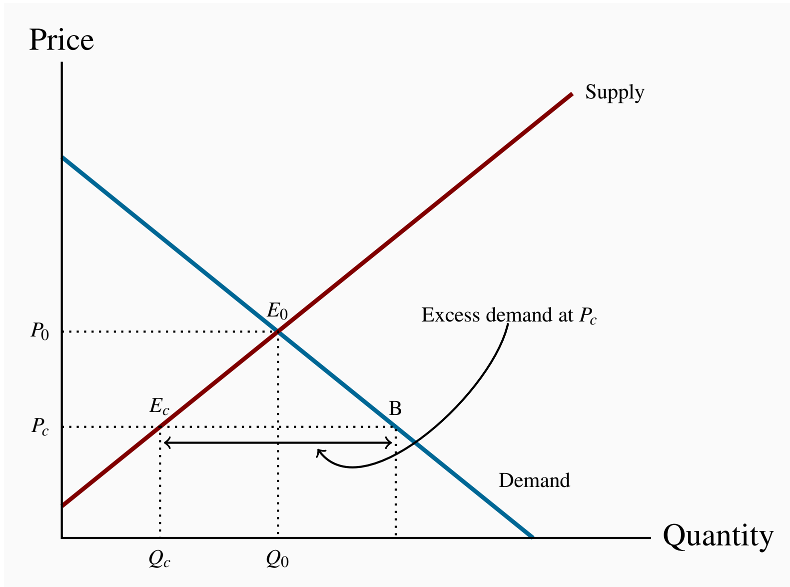
Price ceilings – rental boards

Ceilings mean that suppliers cannot legally charge more than a specific price. Limits on apartment rents are one form of ceiling. In times of emergency – such as flooding or famine, price controls are frequently imposed on foodstuffs, in conjunction with rationing, to ensure that access is not determined by who has the most income. The problem with price ceilings, however, is that they leave demand unsatisfied, and therefore they must be accompanied by some other allocation mechanism.

Consider an environment where, for some reason – perhaps a sudden and unanticipated growth in population – rents increase. Let the resulting equilibrium be defined by the point E_0 in Figure 3.6. If the government were to decide that this is an unfair price because it places hardships on low- and middle-income households, it might impose a price limit, or ceiling, of P_c. The problem with such a limit is that excess demand results: Individuals want to rent more apartments than are available in the city. In a free market the price would adjust upward to eliminate the excess demand, but in this controlled environment it cannot. So some other way of allocating the available supply between demanders must evolve.

In reality, most apartments are allocated to those households already occupying them. But what happens when such a resident household decides to purchase a home or move to another city? In a free market, the landlord could increase the rent in accordance with market pressures. But in a controlled market a city's rental tribunal may restrict the annual rent increase to just a couple of percent and the demand may continue to outstrip supply. So how does the stock of apartments get allocated between the potential renters? One allocation method is well known: The existing tenant informs her friends of her plan to move, and the friends are the first to apply to the landlord to occupy the apartment. But that still leaves much unmet demand. If this is a student rental market, students whose parents live nearby may simply return 'home'. Others may chose to move to a part of the city where rents are more affordable.

Figure 3.6: The effect of a price ceiling



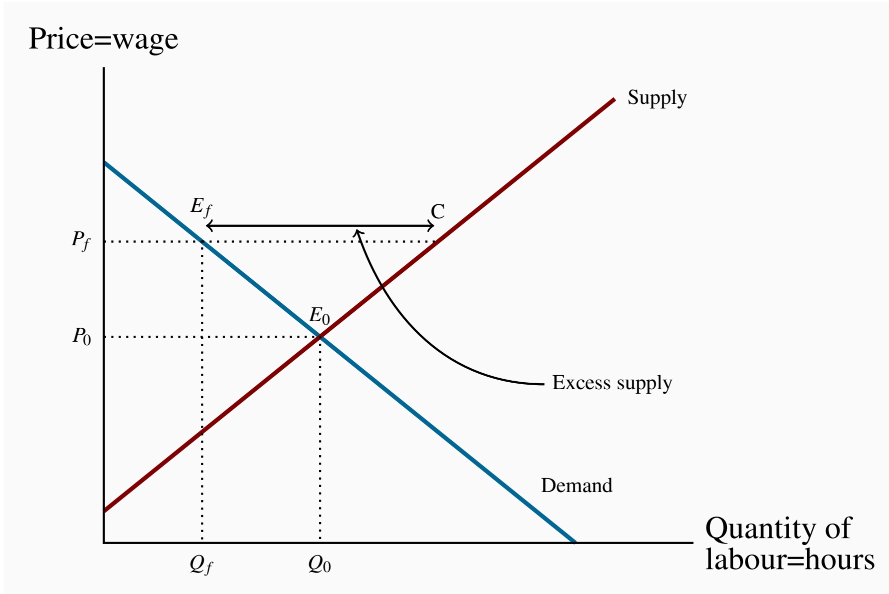
The free market equilibrium occurs at E_0. A price ceiling at P_c holds down the price but leads to excess demand E_cB, because Q_c is the quantity traded. A price ceiling above P_0 is irrelevant since the free market equilibrium E_0 can still be attained.

However, rent controls sometimes yield undesirable outcomes. Rent controls are widely studied in economics, and the consequences are well understood: Landlords tend not to repair or maintain their rental units in good condition if they cannot obtain the rent they believe they are entitled to. Accordingly, the residential rental stock deteriorates. In addition, builders realize that more money is to be made in building condominium units than rental units, or in *converting rental units to condominiums*. The frequent consequence is thus a *reduction* in supply and a reduced quality. Market forces are hard to circumvent because, as we emphasized in Chapter 1, economic players react to the incentives they face. These outcomes are examples of what we call the *law of unintended consequences*.

Price floors – minimum wages

An effective price floor sets the price *above* the market-clearing price. A minimum wage is the most widespread example in the Canadian economy. Provinces each set their own minimum, and it is seen as a way of protecting the well-being of low-skill workers. Such a floor is illustrated in Figure 3.7. The free-market equilibrium is again E_0, but the effective market outcome is the combination of price and quantity corresponding to the point E_f at the price floor, P_f. In this instance, there is excess supply equal to the amount E_fC.

Figure 3.7: Price floor – minimum wage



In a free market the equilibrium is E_0. A minimum wage of P_f raises the hourly wage, but reduces the hours demanded to Q_f. Thus E_fC is the excess supply.

Note that there is a similarity between the outcomes defined in the floor and ceiling cases: The quantity actually traded is *the lesser of the supply quantity and demand quantity at the going price: The short side dominates.*

If price floors, in the form of minimum wages, result in some workers going unemployed, why do governments choose to put them in place? The excess supply in this case corresponds to unemployment – more individuals are willing to work for the going wage than buyers (employers) wish to employ. The answer really depends upon the magnitude of the excess supply. In particular, suppose, in Figure 3.7 that the supply and demand curves going through the equilibrium E_0 were more 'vertical'. This would result in a smaller excess supply than is represented with the existing supply and demand curves. This would mean in practice that a higher wage could go to workers, making them better off, without causing substantial unemployment. This is the trade off that governments face: With a view to increasing the purchasing power of generally lower-skill individuals, a minimum wage is set, hoping that the negative impact on employment will be small. We will return to this in the next chapter, where we examine the responsiveness of supply and demand curves to different prices.

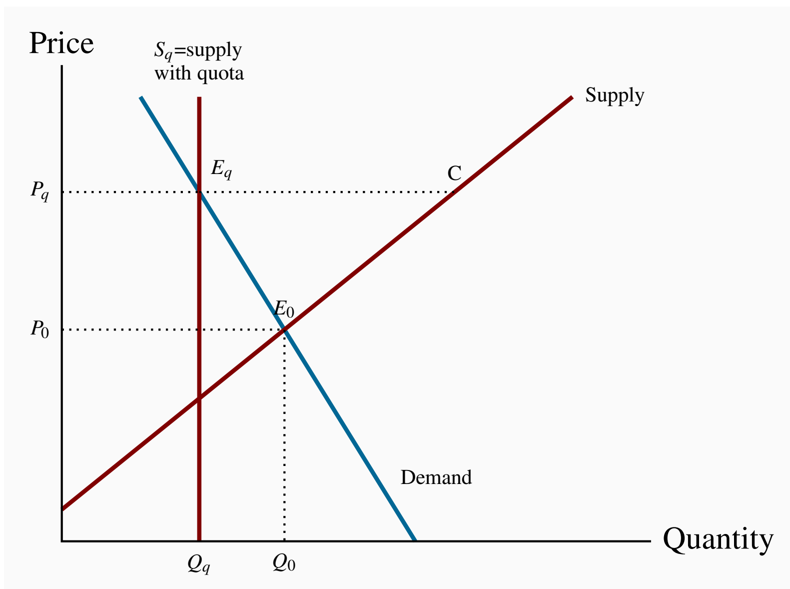
Quotas – agricultural supply

A quota represents the right to supply a specified quantity of a good to the market. It is a means of keeping prices higher than the free-market equilibrium price. As an alternative to imposing a price floor, the government can generate a high price by restricting supply.

Agricultural markets abound with examples. In these markets, farmers can supply only what they are permitted by the quota they hold, and there is usually a market for these quotas. For example, in several Canadian provinces it currently costs in the region of $30,000 to purchase a quota granting the right to sell the milk of one cow. The cost of purchasing quotas can thus easily outstrip the cost of a farm and herd. Canadian cheese importers must pay for the right to import cheese from abroad. Restrictions also apply to poultry. The impact of all of these restrictions is to raise the domestic price above the free market price.

In Figure 3.8, the free-market equilibrium is at E_0. In order to raise the price above P_0, the government restricts supply to Q_q by granting quotas, which permit producers to supply a limited amount of the good in question. This supply is purchased at the price equal to P_q. From the standpoint of farmers, a higher price might be beneficial, even if they get to supply a smaller quantity, provided the amount of revenue they get as a result is as great as the revenue in the free market.

Figure 3.8: The effect of a quota



The government decides that the equilibrium price P_0 is too low. It decides to boost price by reducing supply from Q_0 to Q_q. It achieves this by requiring producers to have a production quota. This is equivalent to fixing supply at S_q.

Marketing boards – milk and maple syrup

A marketing board is a means of insuring that a quota or price floor can be maintained. Quotas are frequent in the agriculture sector of the economy. One example is maple syrup in Quebec. The Federation of Maple Syrup Producers of Quebec has the sole right to market maple syrup. All producers must sell their syrup through this marketing board. The board thus has a particular type of power in the market: it has control of the market at the wholesale end, because it is a sole buyer.

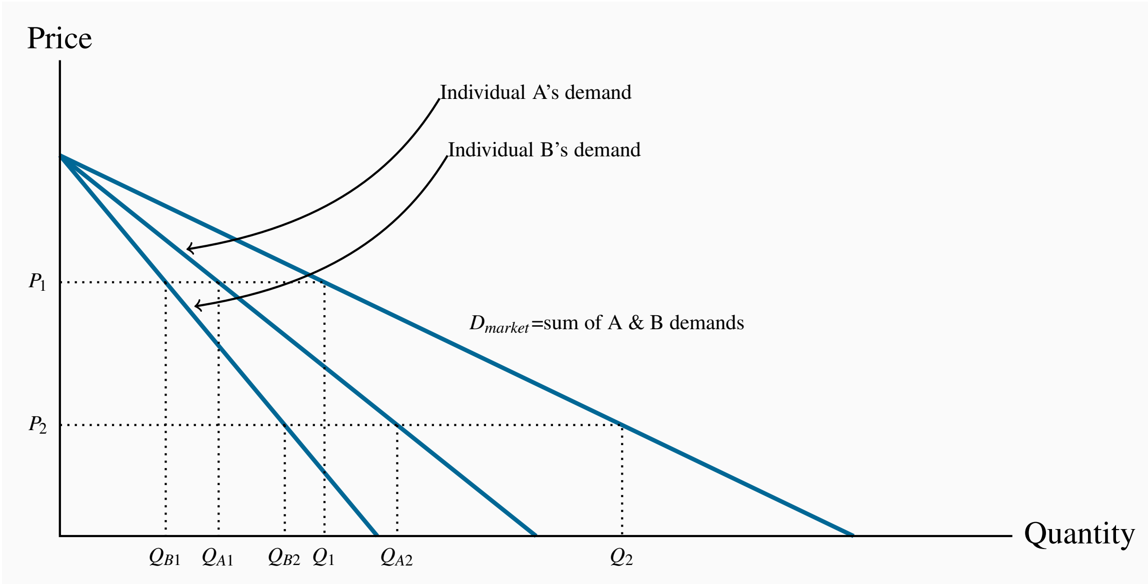
The Federation increases the total revenue going to producers by artificially restricting the supply to the market. The Federation calculates that by reducing supply and selling it at a higher price, more revenue will accrue to the producers. This is illustrated in Figure 3.8. The market equilibrium is given by E_0, but the Federation restricts supply to the quantity Q_q, which is sold to buyers at price P_q. To make this possible the total supply must be restricted; otherwise producers would supply the amount given by the point C on the supply curve, and this would result in excess supply in the amount E_qC. In order to restrict supply to Q_q in total, individual producers are limited in what they can sell to the Federation; they have a quota, which gives them the right to produce and sell no more than a specified amount. This system of quotas is necessary to eliminate the excess supply that would emerge at the above-equilibrium price P_q.

We will return to this topic in Chapter 4. For the moment, to see that this type of revenue-increasing outcome is possible, examine Table 3.1 again. At this equilibrium price of $4 the quantity traded is 6 units, yielding a total expenditure by buyers (revenue to suppliers) of $24. However, if the supply were restricted and a price of $5 were set, the expenditure by buyers (revenue to suppliers) would rise to $25.

3.8 Individual and market functions

Markets are made up of many individual participants on the demand and supply side. The supply and demand functions that we have worked with in this chapter are those for the total of all participants on each side of the market. But how do we arrive at such market functions when the economy is composed of individuals? We can illustrate how, with the help of Figure 3.9.

Figure 3.9: Summing individual demands



At P_1 individual A purchases Q_{A1} and B purchases Q_{B1}. The total demand is the sum of these individual demands at this price (Q_1). At P_2 individual demands are summed to Q_2. Since the points Q_1 and Q_2 define the demands of the market participants it follows that market demand is the horizontal sum of these curves.

To concentrate on the essentials, imagine that there are just two buyers of chocolate cookies in the economy. A has a stronger preference for cookies than B, so his demand is greater. To simplify, let the two demands have the same intercept on the vertical axis. The curves D_A and D_B indicate how many cookies A and B, respectively, will buy at each price. The market demand indicates how much they buy *together* at any price. Accordingly, at P_1, A and B purchase the quantities Q_{A1} and Q_{B1} respectively. Thus $Q_1 = Q_{A1} + Q_{B1}$. At a price P_2, they purchase Q_{A2} and Q_{B2}. Thus $Q_2 = Q_{A2} + Q_{B2}$. The **market demand** is therefore the horizontal sum of the individual demands at these prices. In the figure this is defined by D_{market}.

Market demand: the horizontal sum of individual demands.

3.9 Useful techniques – demand and supply equations

The supply and demand functions, or equations, underlying Table 3.1 and Figure 3.2 can be written in their mathematical form:

$$\text{Demand: } P = 10 - Q$$

$$\text{Supply: } P = 1 + (1/2)Q$$

A straight line is represented completely by the intercept and slope. In particular, if the variable P is on the vertical axis and Q on the horizontal axis, the straight-line equation relating P and Q

is defined by $P = a + bQ$. Where the line is negatively sloped, as in the demand equation, the parameter b must take a negative value. By observing either the data in Table 3.1 or Figure 3.2 it is clear that the vertical intercept, a, takes a value of \$10. The vertical intercept corresponds to a zero-value for the Q variable. Next we can see from Figure 3.2 that the slope (given by the rise over the run) is 10/10 and hence has a value of -1. Accordingly the demand equation takes the form $P = 10 - Q$.

On the supply side the price-axis intercept, from either the figure or the table, is clearly 1. The slope is one half, because a two-unit change in quantity is associated with a one-unit change in price. This is a positive relationship obviously so the supply curve can be written as $P = 1 + (1/2)Q$.

Where the supply and demand curves intersect is the market equilibrium; that is, the price-quantity combination is the same for both supply and demand where the supply curve takes on the same values as the demand curve. This unique price-quantity combination is obtained by equating the two curves: If Demand=Supply, then

$$10 - Q = 1 + (1/2)Q.$$

Gathering the terms involving Q to one side and the numerical terms to the other side of the equation results in $9 = 1.5Q$. This implies that the equilibrium quantity must be 6 units. And this quantity must trade at a price of \$4. That is, when the price is \$4 both the quantity demanded and the quantity supplied take a value of 6 units.

Modelling market interventions using equations

To illustrate the impact of market interventions examined in Section 3.7 on our numerical market model for natural gas, suppose that the government imposes a minimum price of \$6 – above the equilibrium price obviously. We can easily determine the quantity supplied and demanded at such a price. Given the supply equation

$$P = 1 + (1/2)Q,$$

it follows that at $P = 6$ the quantity supplied is 10. This follows by solving the relationship $6 = 1 + (1/2)Q$ for the value of Q. Accordingly, suppliers *would like to supply* 10 units at this price.

Correspondingly on the demand side, given the demand curve

$$P = 10 - Q,$$

with a price given by $P = \$6$, it must be the case that $Q = 4$. So buyers *would like to buy* 4 units at that price: There is excess supply. But we know that the short side of the market will win out, and so the actual amount traded at this restricted price will be 4 units.

Conclusion

We have covered a lot of ground in this chapter. It is intended to open up the vista of economics to the new student in the discipline. Economics is powerful and challenging, and the ideas we have developed here will serve as conceptual foundations for our exploration of the subject. Our next chapter deals with measurement and responsiveness.

Key Terms

Demand is the quantity of a good or service that buyers wish to purchase at each possible price, with all other influences on demand remaining unchanged.

Supply is the quantity of a good or service that sellers are willing to sell at each possible price, with all other influences on supply remaining unchanged.

Quantity demanded defines the amount purchased at a particular price.

Quantity supplied refers to the amount supplied at a particular price.

Equilibrium price: equilibrates the market. It is the price at which quantity demanded equals the quantity supplied.

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

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

Short side of the market determines outcomes at prices other than the equilibrium.

Demand curve is a graphical expression of the relationship between price and quantity demanded, with other influences remaining unchanged.

Supply curve is a graphical expression of the relationship between price and quantity supplied, with other influences remaining unchanged.

Substitute goods: when a price reduction (rise) for a related product reduces (increases) the demand for a primary product, it is a substitute for the primary product.

Complementary goods: when a price reduction (rise) for a related product increases (reduces) the demand for a primary product, it is a complement for the primary product.

Inferior good is one whose demand falls in response to higher incomes.

Normal good is one whose demand increases in response to higher incomes.

Comparative static analysis compares an initial equilibrium with a new equilibrium, where the difference is due to a change in one of the other things that lie behind the demand curve or the supply curve.

Price controls are government rules or laws that inhibit the formation of market-determined prices.

Quotas are physical restrictions on output.

Market demand: the horizontal sum of individual demands.

EXERCISES FOR CHAPTER 3

Exercise 3.1 The supply and demand for concert tickets are given in the table below.

Price ($)	0	4	8	12	16	20	24	28	32	36	40
Quantity demanded	15	14	13	12	11	10	9	8	7	6	5
Quantity supplied	0	0	0	0	0	1	3	5	7	9	11

(a) Plot the supply and demand curves to scale and establish the equilibrium price and quantity.

(b) What is the excess supply or demand when price is $24? When price is $36?

(c) Describe the market adjustments in price induced by these two prices.

(d) *Optional*: The functions underlying the example in the table are linear and can be presented as $P = 18 + 2Q$ (supply) and $P = 60 - 4Q$ (demand). Solve the two equations for the equilibrium price and quantity values.

Exercise 3.2 Illustrate in a supply/demand diagram, by shifting the demand curve appropriately, the effect on the demand for flights between Calgary and Winnipeg as a result of:

(a) Increasing the annual government subsidy to *Via Rail*.

(b) Improving the Trans-Canada highway between the two cities.

(c) The arrival of a new budget airline on the scene.

Exercise 3.3 A new trend in US high schools is the widespread use of chewing tobacco. A recent survey indicates that 15 percent of males in upper grades now use it – a figure not far below the use rate for cigarettes. This development came about in response to the widespread implementation by schools of regulations that forbade cigarette smoking on and around school property. Draw a supply-demand equilibrium for each of the cigarette and chewing tobacco markets before and after the introduction of the regulations.

Exercise 3.4 The following table describes the demand and supply conditions for labour.

Price ($) = wage rate	0	10	20	30	40	50	60	70	80	90	100	110	120	130	140	150	160	170
Quantity demanded	1020	960	900	840	780	720	660	600	540	480	420	360	300	240	180	120	60	0
Quantity supplied	0	0	0	0	0	0	30	60	90	120	150	180	210	240	270	300	330	360

(a) Graph the functions and find the equilibrium price and quantity by equating demand and supply.

(b) Suppose a price ceiling is established by the government at a price of \$120. This price is below the equilibrium price that you have obtained in part (a). Calculate the amount that would be demanded and supplied and then calculate the excess demand.

Exercise 3.5 In Exercise 3.4, suppose that the supply and demand describe an agricultural market rather than a labour market, and the government implements a price floor of \$140. This is greater than the equilibrium price.

(a) Estimate the quantity supplied and the quantity demanded at this price, and calculate the excess supply.

(b) Suppose the government instead chose to maintain a price of \$140 by implementing a system of quotas. What quantity of quotas should the government make available to the suppliers?

Exercise 3.6 In Exercise 3.5, suppose that, at the minimum price, the government buys up all of the supply that is not demanded, and exports it at a price of \$80 per unit. Compute the cost to the government of this operation.

Exercise 3.7 Let us sum two demand curves to obtain a 'market' demand curve. We will suppose there are just two buyers in the market. Each of the individual demand curves has a price intercept of \$42. One has a quantity intercept of 126, the other 84.

(a) Draw the demands either to scale or in an Excel spreadsheet, and label the intercepts on both the price and quantity axes.

(b) Determine how much would be purchased in the market at prices \$10, \$20, and \$30.

(c) *Optional*: Since you know the intercepts of the market (total) demand curve, can you write an equation for it?

Exercise 3.8 In Exercise 3.7 the demand curves had the same price intercept. Suppose instead that the first demand curve has a price intercept of \$36 and a quantity intercept of 126; the other individual has a demand curve defined by a price intercept of \$42 and a quantity intercept of 84. Graph these curves and illustrate the market demand curve.

Exercise 3.9 Here is an example of a demand curve that is not linear:

Price (\$)	4	3	2	1	0
Quantity demanded	25	100	225	400	625

(a) Plot this demand curve to scale or in Excel.

(b) If the supply function in this market is $P = 2$, plot this function in the same diagram.

(c) Determine the equilibrium quantity traded in this market.

Exercise 3.10 The football stadium of the University of the North West Territories has 30 seats. The demand curve for tickets has a price intercept of $36 and a quantity intercept of 72.

(a) Draw the supply and demand curves to scale in a graph or in Excel. (This demand curve has the form $P = 36 - 0.5 \times Q$.)

(b) Determine the equilibrium admission price, and the amount of revenue generated from ticket sales for each game.

(c) A local alumnus and benefactor offers to install 6 more seats at no cost to the University. Compute the price that would be charged with this new supply and compute the revenue that would accrue at this new equilibrium price. Should the University accept the offer to install the seats?

(d) Redo the previous part of this question, assuming that the initial number of seats is 40, and the University has the option to increase capacity to 46 at no cost to itself. Should the University accept the offer in this case?

Exercise 3.11 Suppose farm workers in Mexico are successful in obtaining a substantial wage increase. Illustrate the effect of this on the price of lettuce in the Canadian winter, using a supply and demand diagram, on the assumption that all lettuce in Canada is imported during its winter.

Part Two

Responsiveness and the Value of Markets

The degree to which individuals or firms, or any economic agent, respond to incentives is important to ascertain for pricing and policy purposes: If prices change, to what degree will suppliers and buyers respond? How will markets respond to taxes? Chapter 4 explores and develops the concept of elasticity, which is the word economists use to define responsiveness. A meaningful metric, one formulated in percentage terms, that is applicable to virtually any market or incentive means that behaviours can be compared in different environments.

In Chapter 5 we explore how markets allocate resources and how the well-being of society's members is impacted by uncontrolled and controlled markets. A central theme of this chapter is that markets are very useful environments, but, if they are to serve the social interest, need to be controlled in many circumstances.

Chapter 4

Measures of response: Elasticities

In this chapter we will explore:

4.1 Price responsiveness of demand

Put yourself in the position of an entrepreneur. One of your many challenges is to price your product appropriately. You may be Michael Dell choosing a price for your latest computer, or the local restaurant owner pricing your table d'hôte, or you may be pricing your part-time snow-shoveling service. A key component of the pricing decision is to know how *responsive* your market is to variations in your pricing. How we measure responsiveness is the subject matter of this chapter.

We begin by analyzing the responsiveness of consumers to price changes. For example, consumers tend not to buy much more or much less food in response to changes in the general price level of food. This is because food is a pretty basic item for our existence. In contrast, if the price of new textbooks becomes higher, students may decide to search for a second-hand copy, or make do with lecture notes from their friends or downloads from the course web site. In the latter case students have ready alternatives to the new text book, and so their expenditure patterns can be expected to reflect these options, whereas it is hard to find alternatives to food. In the case of food consumers are not very responsive to price changes; in the case of textbooks they are. The word 'elasticity' that appears in this chapter title is just another term for this concept of responsiveness. Elasticity has many different uses and interpretations, and indeed more than one way of being measured in any given situation. Let us start by developing a suitable numerical measure.

The slope of the demand curve suggests itself as one measure of responsiveness: If we lowered the price of a good by $1, for example, how many more units would we sell? The difficulty with this measure is that it does not serve us well when comparing different products. One dollar may be a substantial part of the price of your morning coffee and croissant, but not very important if buying a computer or tablet. Accordingly, when goods and services are measured in different units (croissants versus tablets), or when their prices are very different, it is often best to use a *percentage* change measure, which is *unit-free*.

The **price elasticity of demand** is measured as the percentage change in quantity demanded, divided by the percentage change in price. Although we introduce several other elasticity measures later, when economists speak of the demand elasticity they invariably mean the price elasticity of demand defined in this way.

> The **price elasticity of demand** is measured as the percentage change in quantity demanded, divided by the percentage change in price.

The price elasticity of demand can be written in different forms. We will use the Greek letter epsilon, ε, as a shorthand symbol, with a subscript d to denote demand, and the capital delta, Δ, to denote a change. Therefore, we can write

$$\text{Price elasticity of demand} = \varepsilon_d = \frac{\text{Percentage change in quantity demanded}}{\text{Percentage change in price}}$$

or, using a shortened expression,

$$\varepsilon_d = \frac{\%\Delta Q}{\%\Delta P} \tag{4.1}$$

Calculating the value of the elasticity is not difficult. If we are told that a 10 percent price increase reduces the quantity demanded by 20 percent, then the elasticity value is $-20\%/10\% = -2$. The negative sign denotes that price and quantity move in opposite directions, but for brevity the negative sign is often omitted.

Consider now the data in Table 4.1 and the accompanying Figure 4.1. These data reflect the demand relation for natural gas that we introduced in Chapter 3. Note first that, when the price and quantity change, we must decide what *reference price and quantity* to use in the percentage change calculation in the definition above. We could use the initial or final price-quantity combination, or an average of the two. Each choice will yield a slightly different numerical value for the elasticity. The best convention is to *use the midpoint of the price values and the corresponding midpoint of the quantity values*. This ensures that the elasticity value is the same regardless of whether we start at the higher price or the lower price. Using the subscript 1 to denote the initial value and 2 the final value:

$$\text{Average quantity } \overline{Q} = (Q_1 + Q_2)/2$$

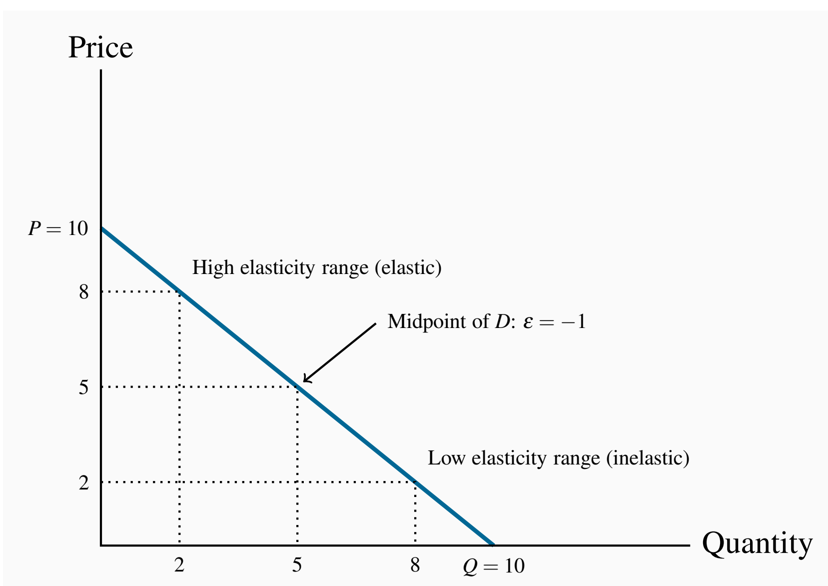
$$\text{Average price } \overline{P} = (P_1 + P_2)/2$$

Table 4.1: The demand for natural gas: Elasticities and revenue

Price ($)	Quantity demanded	Elasticity value	Total revenue ($)
10	0		0
9	1	-9.0	9
8	2		16
7	3	-2.33	21
6	4		24
5	5	-1.0	25
4	6		24
3	7	-0.43	21
2	8		16
1	9	-0.11	9
0	10		0

Elasticity calculations are based upon $2 price changes.

Figure 4.1: Elasticity variation with linear demand



In the high-price region of the demand curve the elasticity takes on a high value. At the midpoint of a linear demand curve the elasticity takes on a value of one, and at lower prices the elasticity value continues to fall.

Using this rule, consider now the value of ε_d when price drops from \$10.00 to \$8.00. The change in price is \$2.00 and the average price is therefore \$9.00 $[= (\$10.00 + \$8.00)/2]$. On the quantity side, demand goes from zero to 2 units (measured in thousands of cubic feet), and the average quantity demanded is therefore $(0+2)/2 = 1$. Putting these numbers into the formula yields:

$$\varepsilon_d = \frac{(Q_2 - Q_1)/\overline{Q}}{(P_2 - P_1)/\overline{P}} = \frac{(2/1)}{-(2/9)} = -\left(\frac{2}{1}\right) \times \left(\frac{9}{2}\right) = -9.$$

Note that the price has declined in this instance and thus the change in price is negative. Continuing down the table in this fashion yields the full set of elasticity values in the third column.

The demand elasticity is said to be *high* if it is a large negative number; the large number denotes a high degree of sensitivity. Conversely, the elasticity is *low* if it is a small negative number. High and low refer to the size of the number, ignoring the negative sign. The term *arc elasticity* is also used to define what we have just measured, indicating that it defines consumer responsiveness over a segment or *arc* of the demand curve.

It is helpful to analyze this numerical example by means of the corresponding demand curve that is plotted in Figure 4.1, and which we used in Chapter 3. It is a straight-line demand curve; but, despite this, the elasticity is not constant. At high prices the elasticity is high; at low prices it is low. The intuition behind this pattern is as follows. When the price is high, a given price change represents a small *percentage* change, because the average price in the price-term denominator is large. At high prices the quantity demanded is small and therefore the percentage quantity change tends to be large due to the small quantity value in its denominator. In sum, at high prices the elasticity is large; it contains a large numerator and a small denominator. By the same reasoning, at low prices the elasticity is small.

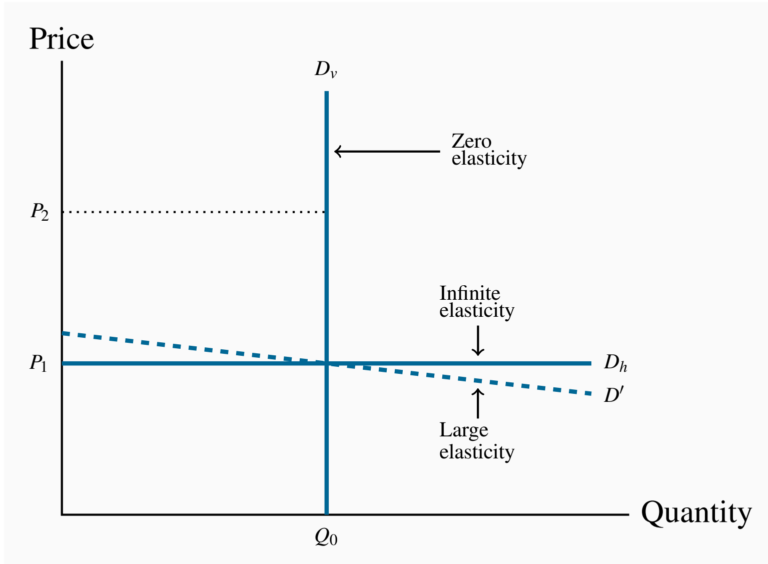
We can carry this reasoning one step further to see what happens when the demand curve intersects the axes. At the horizontal axis the average price is tending towards zero. Since this extremely small value appears in the denominator of the price term it means that the price term as a whole is extremely large. Accordingly, with an extremely large value in the denominator of the elasticity expression, the whole ratio is tending towards a zero value. By the same reasoning the elasticity value at the vertical intercept is tending towards an infinitely large value.

Extreme cases

The elasticity decreases in going from high prices to low prices. This is true for most non-linear demand curves also. Two exceptions are when the demand curve is horizontal and when it is vertical.

When the demand curve is vertical, no quantity change results from a change in price from P_1 to P_2, as illustrated in Figure 4.2 using the demand curve D_v. Therefore, the numerator in Equation 4.1 is zero, and the elasticity has a zero value.

Figure 4.2: Limiting cases of price elasticity

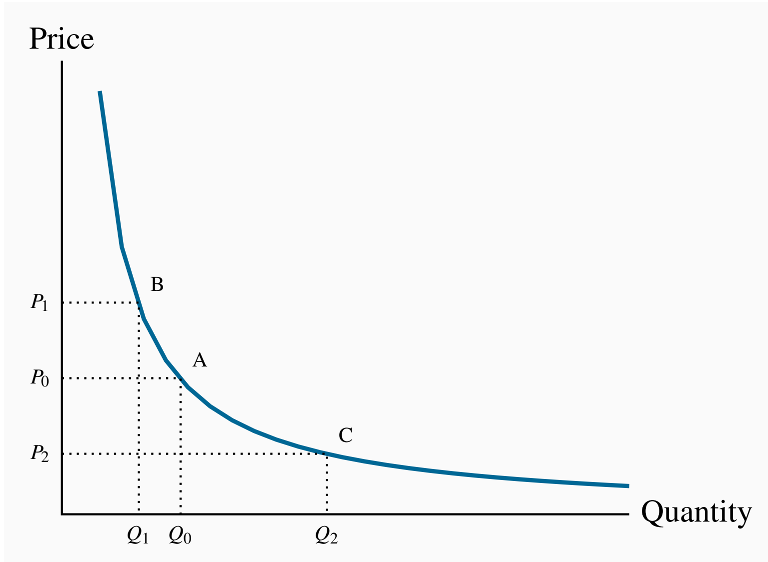


When the demand curve is vertical (D_v), the elasticity is zero: A change in price from P_1 to P_2 has no impact on the quantity demanded because the numerator in the elasticity formula has a zero value. When D becomes more horizontal the elasticity becomes larger and larger at P_1, eventually becoming infinite.

In the horizontal case, we say that the elasticity is *infinite*, which means that any percentage price change brings forth an infinite quantity change! This case is also illustrated in Figure 4.2 using the demand curve D_h. As with the vertical demand curve, this is not immediately obvious. So consider a demand curve that is almost horizontal, such as D' instead of D_h. In this instance, we can achieve large changes in quantity demanded by implementing very small price changes. In terms of Equation 4.1, the numerator is large and the denominator small, giving rise to a large elasticity. Now imagine that this demand curve becomes ever more elastic (horizontal). The same quantity response can be obtained with a smaller price change, and hence the elasticity is larger. Pursuing this idea, we can say that, *as the demand curve becomes ever more elastic, the elasticity value tends towards infinity.*

A *non-linear demand curve* is illustrated in Figure 4.3. If price increases from P_0 to P_1, the corresponding quantity change is given by $(Q_0 - Q_1)$. When the price declines to P_2 the quantity increases from Q_0 to Q_2. When statisticians study data to determine how responsive purchases are to price changes they do not always find a linear relationship between price and quantity. But a linear relationship is frequently a good approximation or representation of actual data and we will continue to analyze responsiveness in a linear framework in this chapter.

Figure 4.3: Non-linear demand curves



When the demand curve is non-linear the slope changes with the price. Hence, equal price changes do not lead to equal quantity changes: The quantity change associated with a change in price from P_0 to P_1 is smaller than the change in quantity associated with the same change in price from P_0 to P_2.

Elastic and inelastic demands

While the elasticity value falls as we move down the demand curve, an important dividing line occurs at the value of -1. This is illustrated in Table 4.1, and is a property of all straight-line demand curves. Disregarding the negative sign, demand is said to be **elastic** if the price elasticity is greater than unity, and **inelastic** if the value lies between unity and 0. It is **unit elastic** if the value is exactly one.

> Demand is **elastic** if the price elasticity is greater than unity. It is **inelastic** if the value lies between unity and 0. It is **unit elastic** if the value is exactly one.

Economists frequently talk of goods as having a "high" or "low" demand elasticity. What does this mean, given that the elasticity varies throughout the length of a demand curve? It signifies that, *at the price usually charged*, the elasticity has a high or low value. For example, your weekly demand for regular coffee at Starbucks might be unresponsive to variations in price around the value of \$3.00, but if the price were \$6, you might be more responsive to price variations. Likewise, when we stated at the beginning of this chapter that the demand for food tends to be inelastic, we really meant that *at the price we customarily face for food*, demand is inelastic.

Determinants of price elasticity

Why is it that the price elasticities for some goods and services are high and for others low?

- One answer lies in *tastes*: If a good or service is a basic necessity in one's life, then price variations have a minimal effect on the quantity demanded, and these products thus have a relatively inelastic demand.
- A second answer lies in the *ease with which we can substitute* alternative goods or services for the product in question. If *Apple* Corporation had no serious competition in the smartphone market, it could price its products even higher than in the presence of *Samsung* and *Google*, who also supply smart phones. A supplier who increases her price will lose more sales if there are ready substitutes to which buyers can switch, than if no such substitutes exist. It follows that a critical role for the marketing department in a firm is to convince buyers of the uniqueness of the firm's product.
- Where *product groups* are concerned, the price elasticity of demand for one product is necessarily higher than for the group as a whole: Suppose the price of one computer tablet brand alone falls. Buyers would be expected to substitute towards this product in large numbers – its manufacturer would find demand to be highly responsive. But if *all* brands are reduced in price, the increase in demand for any one will be more muted. In essence, the one tablet whose price falls has several close substitutes, but tablets in the aggregate do not.
- Finally, there is a *time dimension* to responsiveness, and this is explored in Section 4.3.

4.2 Price elasticities and public policy

In Chapter 3 we explored the implications of putting price floors and supply quotas in place. We saw that price floors can lead to excess supply. An important public policy question therefore is why these policies actually exist. It turns out that we can understand why with the help of elasticity concepts.

Price elasticity and expenditure

Let us return to Table 4.1 and explore what happens to total expenditure/revenue as the price varies. Since total revenue is simply the product of price times quantity it can be computed from the first two columns. The result is given in the final column. We see immediately that total expenditure on the good is highest at the midpoint of the demand curve corresponding to these data. At a price of \$5 expenditure is \$25. No other price yields more expenditure or revenue. Obviously the value \$5 is midway between the zero value and the price or quantity intercept of the demand curve in Figure 4.1. This is a general result for linear demand curves: Expenditure is greatest at the midpoint, and the mid-price corresponds to the mid-quantity on the horizontal axis.

Geometrically this can be seen from Figure 4.1. Since expenditure is the product of price and quantity, in geometric terms it is the area of the rectangle mapped out by any price-quantity combination. For example, at $\{P = \$8, Q = 2\}$ total expenditure is \$16 – the area of the rectangle bounded by these price and quantity values. Following this line of reasoning, if we were to compute the area bounded by a price of \$7 and a corresponding quantity of 3 units we get a larger rectangle – a value of \$21. This example indicates that the largest rectangle occurs at the midpoint of the demand curve. As a general geometric rule this is always the case. Hence we can conclude that the price that generates the greatest expenditure is the midpoint of a linear demand curve.

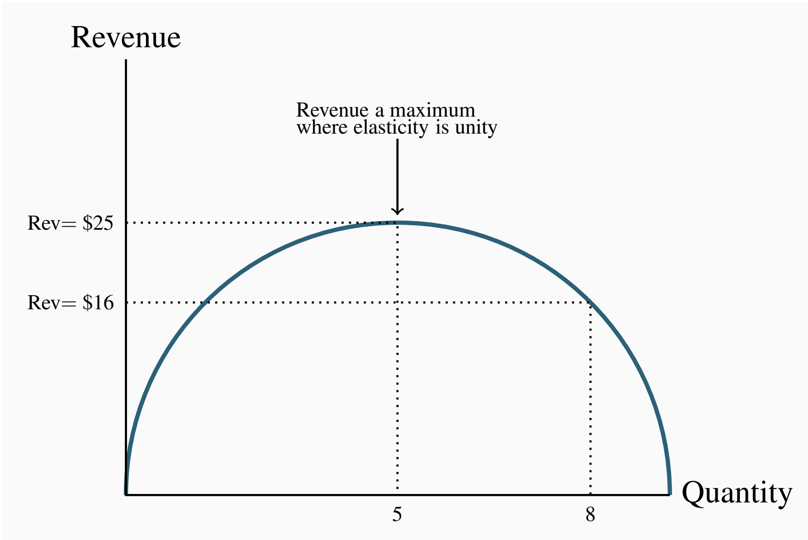
Let us now apply this rule to pricing in the market place. If our existing price is high and our goal is to generate more revenue, then we should reduce the price. Conversely, if our price is low and our goal is again to increase revenue we should raise the price. Starting from a high price let us see why this is so. By lowering the price we induce an increase in quantity demanded. Of course the lower price reduces the revenue obtained on the units already being sold at the initial high price. But since total expenditure increases at the new lower price, it must be the case that *the additional sales caused by the lower price more than compensate for this loss on the units being sold at the initial high price*. But there comes a point when this process ceases. Eventually the loss in revenue on the units being sold at the higher price is not offset by the revenue from additional quantity. We *lose a margin on so many existing units that the additional sales cannot compensate*. Accordingly revenue falls.

Note next that the top part of the demand curve is elastic and the lower part is inelastic. So, as a general rule we can state that:

> *A price decline (quantity increase) on an elastic segment of a demand curve necessarily increases revenue, and a price increase (quantity decline) on an inelastic segment also increases revenue.*

The result is mapped in Figure 4.4, which plots total revenue as a function of the quantity demanded – columns 2 and 4 from Table 4.1. At low quantity values the price is high and the demand is elastic; at high quantity values the price is low and the demand is inelastic. The revenue maximizing point is the midpoint of the demand curve.

Figure 4.4: Total revenue and elasticity



Based upon the data in Table 4.1, revenue increases with quantity sold up to sales of 5 units. Beyond this output, the decline in price that must accompany additional sales causes revenue to decline.

We now have a general conclusion: In order to maximize the possible revenue from the sale of a good or service, it should be priced where the demand elasticity is unity.

Does this conclusion mean that every entrepreneur tries to find this magic region of the demand curve in pricing her product? Not necessarily: Most businesses seek to maximize their *profit* rather than their revenue, and so they have to focus on cost in addition to sales. We will examine this interaction in later chapters. Secondly, not every firm has control over the price they charge; the price corresponding to the unit elasticity may be too high relative to their competitors' price choices. Nonetheless, many firms, especially in the early phase of their life-cycle, focus on revenue growth rather than profit, and so, if they have any power over their price, the choice of the unit-elastic price may be appropriate.

The agriculture problem

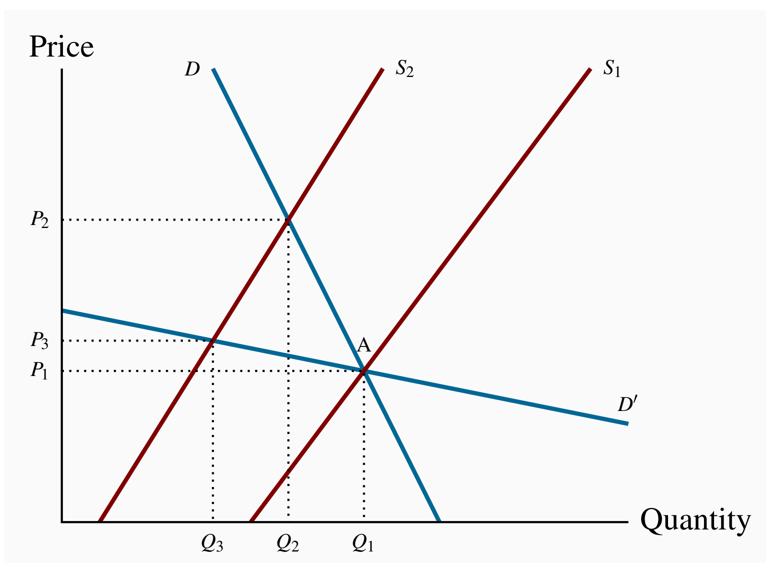
We are now in a position to address the question we posed above: Why are price floors frequently found in agricultural markets? The answer is that governments believe that the pressures of competition would force farm/food prices so low that many farmers would not be able to earn a reasonable income from farming. Accordingly, governments impose price floors. Keep in mind that price floors are prices above the market equilibrium and therefore lead to excess supply.

Since the demand for foodstuffs is inelastic we know that a higher price will induce more revenue, even with a lower quantity being sold. The government can force this outcome on the market by a policy of supply management. It can force farmers in the aggregate to bring only a specific amount of product to the market, and thus ensure that the price floor does not lead to excess supply. This is the system of supply management we observe in dairy markets in Canada, for example, and that we examined in the case of maple syrup in Chapter 3. Its supporters praise it because it helps farmers, its critics point out that higher food prices hurt lower-income households more than high-income households, and therefore it is not a good policy.

Elasticity values are frequently more informative than diagrams and figures. Our natural inclination is to view demand curves with a somewhat vertical profile as being inelastic, and demand curves with a flatter profile as elastic. But we must keep in mind that, as explained in Chapter 2, the vertical and horizontal axis of any diagram can be scaled in such a way as to change the visual impact of the data underlying the curves. But a numerical elasticity value will never deceive in this way. If its value is less than unity it is inelastic, regardless of the visual aspect of the demand curve.

At the same time, if we have two demand curves intersecting at a particular price-quantity combination, we can say that the curve with the more vertical profile is *relatively* more elastic, or less inelastic. This is illustrated in Figure 4.5. It is clear that, *at the price-quantity combination where they intersect*, the demand curve D' will yield a greater (percentage) quantity change than the demand curve D, for a given (percentage) price change. Hence, on the basis of diagrams, we can compare demand elasticities *in relative terms at a point where the two intersect*.

Figure 4.5: The impact of elasticity on quantity fluctuations



In the lower part of the demand curve D, demand is inelastic: At the point A, a shift in supply from S_1 to S_2 induces a large percentage increase in price, and a small percentage decrease in quantity demanded. In contrast, for the demand curve D' that goes through the original equilibrium, the region A is now an *elastic* region, and the impact of the supply shift is contrary: The $\%\Delta P$ is smaller and the $\%\Delta Q$ is larger.

4.3 The time horizon and inflation

The price elasticity of demand is frequently lower in the short run than in the long run. For example, a rise in the price of home heating oil may ultimately induce consumers to switch to natural gas or electricity, but such a transition may require a considerable amount of time. Time is required for decision-making and investment in new heating equipment. A further example is the elasticity of demand for tobacco. Some adults who smoke may be seriously dependent and find quitting almost impossible. Higher prices may provide a stronger incentive to reduce or quit, but successful quitters usually require several attempts before being successful. Several years may be required for the impact of a price increase to be fully apparent. Accordingly when we talk of the short run and the long run, there is no simple rule for defining how long the long run actually is in terms of months or years. In some cases, adjustment may be complete in weeks, in other cases years.

In Chapter 2 we distinguished between real and nominal variables. The former adjust for inflation; the latter do not. Suppose all nominal variables double in value: Every good and service costs twice as much, wage rates double, dividends and rent double, etc. This implies that whatever bundle of goods was previously affordable is still affordable. Nothing has really changed. Demand behaviour is unaltered by this doubling of all prices and all incomes.

How do we reconcile this with the idea that own-price elasticities measure changes in quantity demanded as prices change? Keep in mind that elasticities measure the impact of changing one variable alone, *holding constant all of the others*. But when all variables are changing simultaneously, it is incorrect to think that the impact on quantity of one price or income change is a true measure of responsiveness or elasticity. The *price changes that go into measuring elasticities are therefore changes in prices relative to inflation*.

4.4 Cross-price elasticities – cable or satellite

The price elasticity of demand tells us about consumer responses to price changes in different regions of the demand curve, holding constant all other influences. One of those influences is the price of other goods and services. A **cross-price elasticity** indicates how demand is influenced by changes in the prices of other products.

> The **cross-price elasticity of demand** is the percentage change in the quantity demanded of a product divided by the percentage change in the price of another.

We write the cross price elasticity of the demand for x due to a change in the price of y as

$$\varepsilon_{d(x,y)} = \frac{\text{percentage change in quantity demanded of x}}{\text{percentage change in price of good y}} = \frac{\%\Delta Q_x}{\%\Delta P_y}$$

For example, if the price of cable-supply internet services declines, by how much will the demand for satellite-supply services change? The cross-price elasticity may be positive or negative. These particular goods are clearly *substitutable*, and this is reflected in a *positive* value of this cross-price elasticity: The percentage change in satellite subscribers will be negative in response to a decline in the price of cable; a negative divided by a negative is positive. In contrast, a change in the price of tablets or electronic readers should induce an opposing change in the quantity of e-books purchased: Lower tablet prices will induce greater e-book purchases. In this case the price and quantity movements are in opposite directions and the elasticity is therefore negative – the goods are complements.

Application Box 4.1: Cross-price elasticity of demand between legal and illegal marijuana

In November 2016 Canada's Parliamentary Budget Office produced a research paper on the challenges associated with pricing legalized marijuana. They proposed that taxes should be low rather than high on this product, surprising many health advocates. Specifically they argued that the legal price of marijuana should be just fractionally higher than the price in the illegal market. Otherwise marijuana users would avail of the illegal market supply, which is widely available and of high quality. Effectively their research pointed to a very high cross-price elasticity of demand. This recommendation may mean that tax revenue from marijuana sales will be small, but the size of the illegal market will decline substantially, thereby attaining a prime objective of legalization.

4.5 The income elasticity of demand

In Chapter 3 we stated that higher incomes tend to increase the quantity demanded at any price. To measure the responsiveness of demand to income changes, a unit-free measure exists: The income elasticity of demand. The **income elasticity of demand** is the percentage change in quantity demanded divided by a percentage change in income.

> The **income elasticity of demand** is the percentage change in quantity demanded divided by a percentage change in income.

Let us use the Greek letter eta, η, to define the income elasticity of demand and I to denote income. Then,

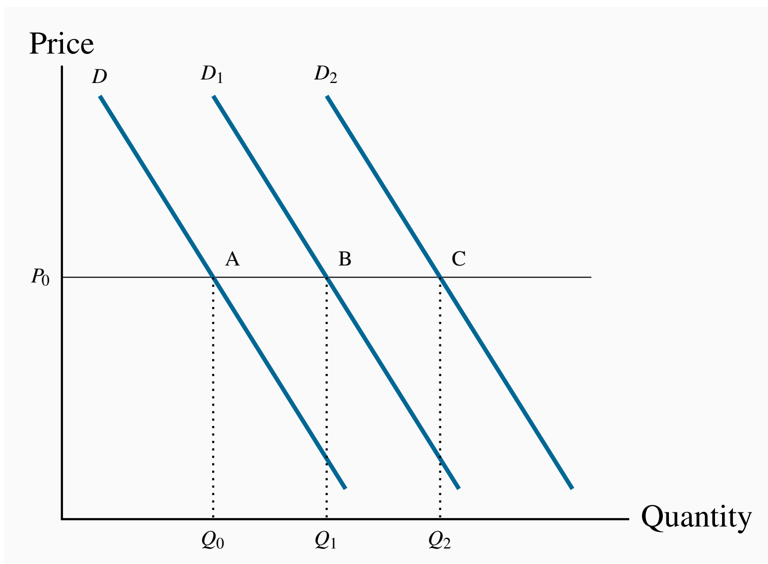
$$\eta_d = \frac{\text{percentage change in quantity demanded}}{\text{percentage change in income}} = \frac{\%\Delta Q}{\%\Delta I}$$

As an example, if monthly income increases by 10 percent, and the quantity of magazines purchased increases by 15 percent, then the income elasticity of demand for magazines is 1.5 in value $(= 15\%/10\%)$. The income elasticity is generally positive, but not always – let us see why.

Normal, inferior, necessary, and luxury goods

The income elasticity of demand, in diagrammatic terms, is a percentage measure of how far the demand curve shifts in response to a change in income. Figure 4.6 shows two possible shifts. Suppose the demand curve is initially the one defined by D, and then income increases. In this example the supply curve is horizontal at the price P_0. If the demand curve shifts to D_1 as a result, the change in quantity demanded at the existing price is $(Q_1 - Q_0)$. However, if instead the demand curve shifts to D_2, that shift denotes a larger change in quantity $(Q_2 - Q_0)$. Since the shift in demand denoted by D_2 exceeds the shift to D_1, the D_2 shift is more responsive to income, and therefore implies a higher income elasticity.

Figure 4.6: Income elasticity and shifts in demand



At the price P_0, the income elasticity measures the percentage horizontal shift in demand caused by some percentage income increase. A shift from A to B reflects a lower income elasticity than a shift to C. A leftward shift in the demand curve in response to an income increase would denote a negative income elasticity – an inferior good.

In this example, the good is a *normal good*, as defined in Chapter 3, because the demand for it increases in response to income increases. If the demand curve were to shift back to the left in response to an increase in income, then the income elasticity would be negative. In such cases the goods or services are *inferior*, as defined in Chapter 3.

Finally, we distinguish between luxuries and necessities. A **luxury** good or service is one whose income elasticity equals or exceeds unity. A **necessity** is one whose income elasticity is greater than zero but less than unity. If quantity demanded is so responsive to an income increase that the percentage increase in quantity demanded exceeds the percentage increase in income, then the elasticity value is in excess of 1, and the good or service is called a luxury. In contrast, if the percentage change in quantity demanded is less than the percentage increase in income, the value is less than unity, and we call the good or service a necessity.

> A **luxury** good or service is one whose income elasticity equals or exceeds unity.
>
> A **necessity** is one whose income elasticity is greater than zero and less than unity.

Luxuries and necessities can also be defined in terms of their share of a typical budget. An income elasticity greater than unity means that the share of an individual's budget being allocated to the product is increasing. In contrast, if the elasticity is less than unity, the budget share is falling. This makes intuitive sense—luxury cars are luxury goods by this definition because they take up a larger share of the incomes of the rich than the non-rich.

Inferior goods are those for which there exist higher-quality, more expensive, substitutes. For example, lower-income households tend to satisfy their travel needs by using public transit. As income rises, households may reduce their reliance on public transit in favour of automobile use (despite the congestion and environmental impacts). Likewise, laundromats are inferior goods in the sense that, as income increases, individuals tend to purchase their own appliances and therefore use laundromat services less. Inferior goods, therefore, have a negative income elasticity: In the income elasticity equation definition, the numerator has a sign opposite to that of the denominator.

> **Inferior goods** have negative income elasticity.

Empirical research indicates that goods like food and fuel have income elasticities less than 1; durable goods and services have elasticities slightly greater than 1; leisure goods and foreign holidays have elasticities very much greater than 1.

Income elasticities are useful in forecasting the demand for particular services and goods in a growing economy. Suppose real income is forecast to grow by 15% over the next five years. If we know that the income elasticity of demand for smart phones is 2.0, we could estimate the anticipated growth in demand by using the income elasticity formula: Since in this case $\eta = 2.0$ and $\%\Delta I = 15\%$ it follows that $2.0 = \%\Delta Q/15\%$. Therefore the predicted demand change must be 30%.

4.6 Elasticity of supply

Now that we have developed the various dimensions of elasticity on the demand side, the analysis of elasticities on the supply side is straightforward. The **elasticity of supply** measures the responsiveness of the quantity supplied to a change in the price.

The **elasticity of supply** measures the responsiveness of quantity supplied to a change in the price.

$$\varepsilon_s = \frac{\text{percentage change in quantity supplied}}{\text{percentage change in price}} = \frac{\%\Delta Q}{\%\Delta P}$$

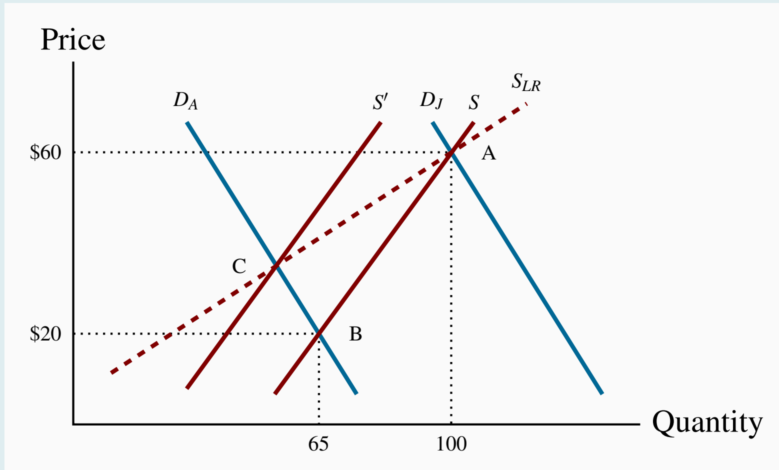
The subscript s denotes supply. This is exactly the same formula as for the demand curve, except that the quantities now come from a supply curve. Furthermore, and in contrast to the demand elasticity, the supply elasticity is generally a positive value because of the positive relationship between price and quantity supplied. The more elastic, or the more responsive, is supply to a given price change, the larger will be the elasticity value. In diagrammatic terms, this means that "flatter" supply curves have a greater elasticity than more "vertical" curves at a given price and quantity combination. Numerically the flatter curve has a larger value than the more vertical supply – try drawing a supply diagram similar to Figure 4.2. Technically, a completely vertical supply curve has a zero elasticity and a horizontal supply curve has an infinite elasticity – just as in the demand cases.

As always we keep in mind the danger of interpreting too much about the value of this elasticity from looking at the visual profiles of supply curves.

Application Box 4.2: The price of oil and the coronavirus

In January 2020, the price of oil in the US traded at \$60 per barrel. The coronavirus then struck the world economy, transport declined dramatically, and by the beginning of April the price had dropped to \$20 per barrel. The quantity of oil traded on world markets fell from approximately 100 million barrels per day in January to 65 million barrels by the start of April, a drop of 35% relative to its January level. This scenario is displayed in Figure 4.7 below.

Figure 4.7: Supply elasticity in the short run and the long run; the oil market in 2020



Demand drops suddenly between January and April. Equilibrium moves from point A to B. In the long run some producers exit and the supply curve shifts towards the origin. Following this, the equilibrium is C. Joining points such as A and C yields a long run supply curve; it is more elastic than the short run supply, when the number of suppliers is fixed.

The January equilibrium is at the point A, and the April equilibrium at point B. The move from A to B was caused by a collapse in demand as illustrated by the shift in the demand curve. We can compute the supply elasticity readily from this example. Note that it was demand that shifted rather than supply so we are observing two points on the supply curve. The supply elasticity, using arc values, is given by (35/82.5)/($40/$40) = 0.42. So the supply curve is inelastic.

Can all oil producers survive in this bear market? The answer is no. It is inexpensive to pump oil in Saudi Arabia, but more costly to produce it from shale or tar sands, and thus some producers are not covering their costs with the price at $20 per barrel. They do not want to shut their operations because closing down and reopening is expensive. They hang on in the hope that the price will return towards $60. If demand does not recover, some suppliers exit the industry with the passage of time. With fewer suppliers the supply curve shifts towards the origin and ultimately another equilibrium price and quantity are established. Call this new equilibrium point C.

The supply elasticity takes on a different value in the short run than in the long run. Supply is more inelastic in the short run. A line through the points A and C would represent the long-run supply curve for the industry. It must be more elastic than the short run supply because the industry has had time to adjust - it is more flexible with the passage of time.

4.7 Elasticities and tax incidence

Elasticity values are critical in determining the impact of a government's taxation policies. The spending and taxing activities of the government influence the use of the economy's resources. By taxing cigarettes, alcohol and fuel, the government can restrict their use; by taxing income, the government influences the amount of time people choose to work. Taxes have a major impact on almost every sector of the Canadian economy.

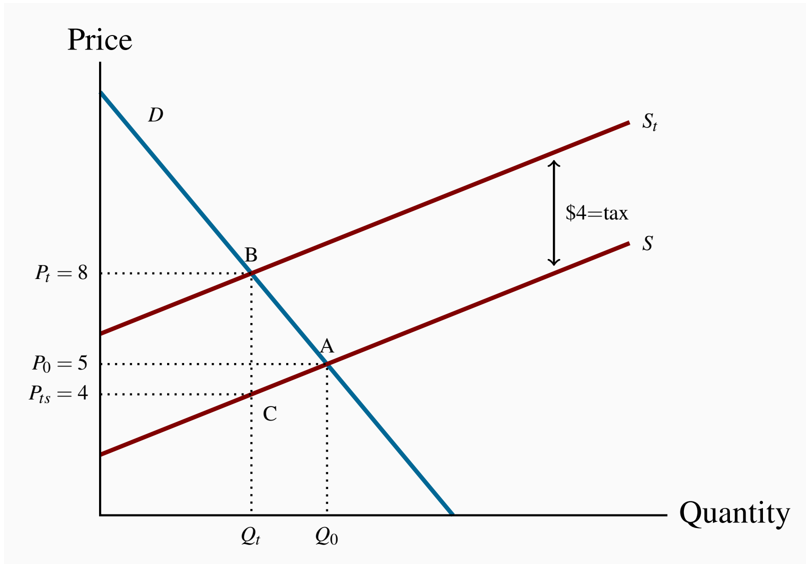
To illustrate the role played by demand and supply elasticities in tax analysis, we take the example of a sales tax. These can be of the *specific* or *ad valorem* type. A *specific* tax involves a fixed dollar levy per unit of a good sold (e.g., $10 per airport departure). An *ad valorem* tax is a percentage levy, such as Canada's Goods and Services tax (e.g., 5 percent on top of the retail price of goods and services). The impact of each type of tax is similar, and we will use the specific tax in our example below.

A layperson's view of a sales tax is that the tax is borne by the consumer. That is to say, if no sales tax were imposed on the good or service in question, the price paid by the consumer would be the same net of tax price as exists when the tax is in place. Interestingly, this is not always the case. The study of the **incidence of taxes** is the study of who really bears the tax burden, and this in turn depends upon supply and demand elasticities.

> **Tax Incidence** describes how the burden of a tax is shared between buyer and seller.

Consider Figures 4.8 and 4.9, which define an imaginary market for inexpensive wine. Let us suppose that, without a tax, the equilibrium price of a bottle of wine is $5, and Q_0 is the equilibrium quantity traded. The pre-tax equilibrium is at the point A. The government now imposes a specific tax of $4 per bottle. The impact of the tax is represented by an upward shift in supply of $4: Regardless of the price that the consumer pays, $4 of that price must be remitted to the government. As a consequence, the price paid to the supplier must be $4 less than the consumer price, and this is represented by twin supply curves: One defines the price at which the supplier is willing to supply (S), and the other is the tax-inclusive supply curve that the consumer faces (S_t).

Figure 4.8: Tax incidence with elastic supply



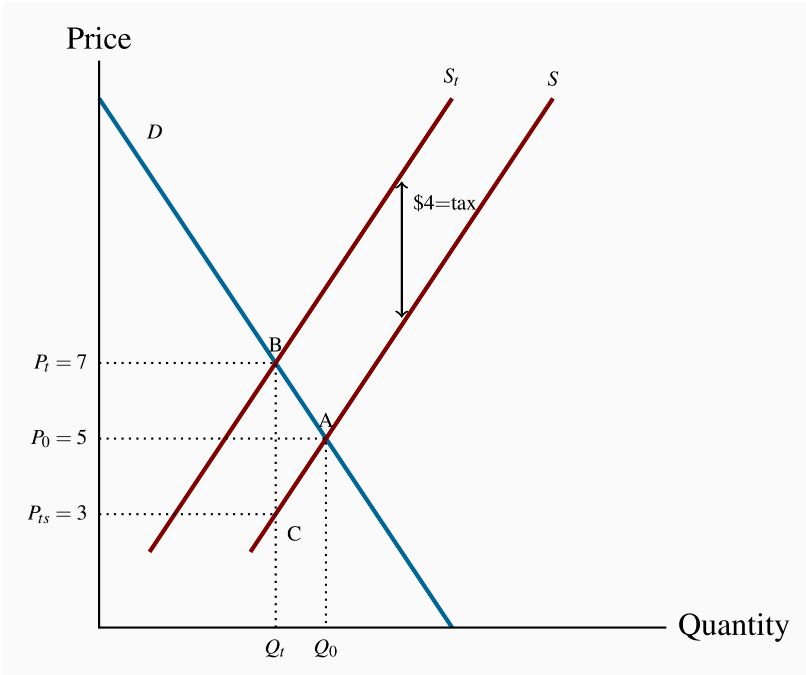
The imposition of a specific tax of $4 shifts the supply curve vertically by $4. The final price at B (P_t) increases by $3 over the equilibrium price at A. At the new quantity traded, Q_t, the supplier gets $4 per unit ($P_{ts}$), the government gets $4 also and the consumer pays $8. The greater part of the incidence is upon the buyer, on account of the relatively elastic supply curve: His price increases by $3 of the $4 tax.

The introduction of the tax in Figure 4.8 means that consumers now face the supply curve S_t. The new equilibrium is at point B. Note that the price has increased by less than the full amount of the tax—in this example it has increased by $3. This is because the reduced quantity at B is provided at a lower supply price: The supplier is willing to supply the quantity Q_t at a price defined by C ($4), which is lower than the price at A ($5).

So what is the incidence of the $4 tax? Since the market price has increased from $5 to $8, and the price obtained by the supplier has fallen by $1, we say that the incidence of the tax falls mainly on the consumer: The price to the consumer has risen by three dollars and the price received by the supplier has fallen by just one dollar.

Consider now Figure 4.9, where the supply curve is less elastic, and the demand curve is unchanged. Again the supply curve must shift upward with the imposition of the $4 specific tax. But here the price received by the supplier is lower than in Figure 4.8, and the price paid by the consumer does not rise as much – the incidence is different. The consumer faces a price increase that is one-half, rather than three-quarters, of the tax value. The supplier faces a lower supply price, and bears a higher share of the tax.

Figure 4.9: Tax incidence with inelastic supply



The imposition of a specific tax of \$4 shifts the supply curve vertically by \$4. The final price at B (P_t) increases by \$2 over the no-tax price at A. At the new quantity traded, Q_t, the supplier gets \$3 per unit (P_{ts}), the government gets \$4 also and the consumer pays \$7. The incidence is shared equally by suppliers and demanders.

We can conclude from this example that, for any given demand, *the more elastic is supply, the greater is the price increase* in response to a given tax. Furthermore, a *more elastic supply curve* means that the *incidence falls more on the consumer*; while a *less elastic supply* curve means the *incidence falls more on the supplier*. This conclusion can be verified by drawing a third version of Figure 4.8 and 4.9, in which the supply curve is horizontal – perfectly elastic. When the tax is imposed the price to the consumer increases by the full value of the tax, and the full incidence falls on the buyer. While this case corresponds to the layperson's intuition of the incidence of a tax, economists recognize it as a special case of the more general outcome, where the incidence falls on both the supply side and the demand side.

These are key results in the theory of taxation. It is equally the case that *the incidence of the tax depends upon the demand elasticity*. In Figure 4.8 and 4.9 we used the same demand curve. However, it is not difficult to see that, if we were to redo the exercise with a demand curve of a different elasticity, the incidence would not be identical. At the same time, the general result on supply elasticities still holds. We will return to this material in Chapter 5.

Statutory incidence

In the above example the tax is analyzed by means of shifting the supply curve. This implies that the supplier is obliged to charge the consumer a tax and then return this tax revenue to the government. But suppose the supplier did not bear the obligation to collect the revenue; instead the buyer is required to send the tax revenue to the government, as in the case of employers who are required to deduct income tax from their employees' pay packages (the employers here are the demanders). If this were the case we could analyze the impact of the tax by reducing the market *demand* curve by the \$4. This is because the demand curve reflects what buyers are willing to pay, and when suppliers are paid in the presence of the tax they will be paid the buyers' demand price minus the tax that the buyers must pay. It is not difficult to show that whether we move the supply curve upward (to reflect the responsibility of the supplier to pay the government) or move the demand curve downward, the outcome is the same – in the sense that the same price and quantity will be traded in each case. Furthermore the incidence of the tax, measured by how the price change is apportioned between the buyers and sellers is also unchanged.

Tax revenues and tax rates

It is useful to relate elasticity values to the policy question of the impact of higher or lower taxes on government tax revenue. Consider a situation in which a tax is already in place and the government considers increasing the rate of tax. Can an understanding of elasticities inform us on the likely outcome? The answer is yes. Suppose that at the initial tax-inclusive price demand is inelastic. We know immediately that a tax rate increase that increases the price must increase total expenditure. Hence the outcome is that the government will get a higher share of an increased total expenditure. In contrast, if demand is elastic at the initial tax-inclusive price a tax rate increase that leads to a higher price will *decrease* total expenditure. In this case the government will get a larger share of a smaller pie – not as valuable from a tax-revenue standpoint as a larger share of a larger pie.

4.8 Technical tricks with elasticities

We can easily compute elasticities at any *point* on a demand curve, rather than over a range or arc, by using the explicit formula for the demand curve. To see this note that we can rewrite Equation 4.1 as:

$$\varepsilon_d = \frac{\%\Delta Q}{\%\Delta P} = \frac{\Delta Q/Q}{\Delta P/P} = \frac{\Delta Q}{\Delta P} \times \frac{P}{Q}.$$

The first term in the final form of this expression $(\Delta Q/\Delta P)$ is obtained from the slope of the demand curve, and the second term (P/Q) is defined by the point on the curve that interests us. For example if our demand curve is $P = 10 - 1 \times Q$, then $\Delta P/\Delta Q = -1$. Inverting this to get $\Delta Q/\Delta P$ yields -1 also. So, the point elasticity value at $\{P = \$6, Q = 4\}$ is $-1 \times 6/4 = -1.5$. This formula provides the elasticity value at a particular point on the demand curve, rather than over a range of values or an arc. Consequently it is called the point elasticity of demand. And obviously we could apply it to a demand curve that is not linear provided we know the mathematical form and are able to establish the slope.

Key Terms

Price elasticity of demand is measured as the percentage change in quantity demanded, divided by the percentage change in price.

Demand is elastic if the price elasticity is greater than unity. It is **inelastic** if the value lies between unity and 0. It is **unit elastic** if the value is exactly one.

Cross-price elasticity of demand is the percentage change in the quantity demanded of a product divided by the percentage change in the price of another.

Income elasticity of demand is the percentage change in quantity demanded divided by a percentage change in income.

Luxury good or service is one whose income elasticity equals or exceeds unity.

Necessity is one whose income elasticity is greater than zero and is less than unity.

Inferior goods have a negative income elasticity.

Elasticity of supply is defined as the percentage change in quantity supplied divided by the percentage change in price.

Tax Incidence describes how the burden of a tax is shared between buyer and seller.

Exercises for Chapter 4

Exercise 4.1 Consider the information in the table below that describes the demand for movie rentals from your on-line supplier Instant Flicks.

Price per movie ($)	Quantity demanded	Total revenue	Elasticity of demand
2	1200		
3	1100		
4	1000		
5	900		
6	800		
7	700		
8	600		

(a) Either on graph paper or a spreadsheet, map out the demand curve.

(b) In column 3, insert the total revenue generated at each price.

(c) At what price is total revenue maximized?

(d) In column 4, compute the elasticity of demand corresponding to each $1 price reduction, using the average price and quantity at each state.

(e) Do you see a connection between your answers in parts (c) and (d)?

Exercise 4.2 Your fruit stall has 100 ripe bananas that must be sold today. Your supply curve is therefore vertical. From past experience, you know that these 100 bananas will all be sold if the price is set at 40 cents per unit.

(a) Draw a supply and demand diagram illustrating the market equilibrium price and quantity.

(b) The demand elasticity is -0.5 at the equilibrium price. But you now discover that 10 of your bananas are rotten and cannot be sold. Draw the new supply curve and calculate the percentage price increase that will be associate with the new equilibrium, on the basis of your knowledge of the demand elasticity.

Exercise 4.3 University fees in the State of Nirvana have been frozen in real terms for 10 years. During this period enrolments increased by 20 percent, reflecting an increase in demand. This means the supply curve is horizontal at a given price.

(a) Draw a supply curve and two demand curves to represent the two equilibria described.

(b) Can you estimate a price elasticity of demand for university education in this market?

(c) In contrast, during the same time period fees in a neighbouring state (where supply is also horizontal) increased by 60 percent and enrolments increased by 15 percent. Illustrate this situation in a diagram, where supply is again horizontal.

Exercise 4.4 Consider the demand curve defined by the information in the table below.

Price of movies	Quantity demanded	Total revenue	Elasticity of demand
2	200		
3	150		
4	120		
5	100		

(a) Plot the demand curve to scale and note that it is non-linear.

(b) Compute the total revenue at each price.

(c) Compute the arc elasticity of demand for each of the three price segments.

Exercise 4.5 Waterson Power Corporation's regulator has just allowed a rate increase from 9 to 11 cents per kilowatt hour of electricity. The short-run demand elasticity is -0.6 and the long-run demand elasticity is -1.2 at the current price..

(a) What will be the percentage reduction in power demanded in the short run (use the midpoint 'arc' elasticity formula)?

(b) What will be the percentage reduction in power demanded in the long run?

(c) Will revenues increase or decrease in the short and long runs?

Exercise 4.6 Consider the own- and cross-price elasticity data in the table below.

		% change in price		
		CDs	**Magazines**	**Cappuccinos**
% change in quantity	**CDs**	-0.25	0.06	0.01
	Magazines	-0.13	-1.20	0.27
	Cappuccinos	0.07	0.41	-0.85

(a) For which of the goods is demand elastic and for which is it inelastic?

(b) What is the effect of an increase in the price of CDs on the purchase of magazines and cappuccinos? What does this suggest about the relationship between CDs and these other commodities; are they substitutes or complements?

(c) In graphical terms, if the price of CDs or the price of cappuccinos increases, illustrate how the demand curve for magazines shifts.

Exercise 4.7 You are responsible for running the Speedy Bus Company and have information about the elasticity of demand for bus travel: The own-price elasticity is -1.4 at the current price. A friend who works in the competing railway company also tells you that she has estimated the cross-price elasticity of train-travel demand with respect to the price of bus travel to be 1.7.

(a) As an economic analyst, would you advocate an increase or decrease in the price of bus tickets if you wished to increase revenue for Speedy?

(b) Would your price decision have any impact on train ridership?

Exercise 4.8 A household's income and restaurant visits are observed at different points in time. The table below describes the pattern.

Income ($)	Restaurant visits	Income elasticity of demand
16,000	10	
24,000	15	
32,000	18	
40,000	20	
48,000	22	
56,000	23	
64,000	24	

(a) Construct a scatter diagram showing quantity on the vertical axis and income on the horizontal axis.

(b) Is there a positive or negative relationship between these variables?

(c) Compute the income elasticity for each income increase, using midpoint values.

(d) Are restaurant meals a normal or inferior good?

Exercise 4.9 The demand for bags of candy is given by $P = 48 - 0.2Q$, and the supply by $P = Q$. The demand intercepts here are $P = \$48$ and $Q = 240$; the supply curve is a 45 degree straight line through the origin.

(a) Illustrate the resulting market equilibrium in a diagram knowing that the demand intercepts are $\{\$48, 240\}$, and that the supply curve is a 45 degree line through the origin.

(b) If the government now puts a \$12 tax on all such candy bags, illustrate on a diagram how the supply curve will change.

(c) Instead of the specific tax imposed in part (b), a percentage tax (ad valorem) equal to 30 percent is imposed. Illustrate how the supply curve would change.

Exercise 4.10 *Optional*: Consider the demand curve $P = 100 - 2Q$. The supply curve is given by $P = 30$.

(a) Draw the supply and demand curves to scale, knowing that the demand curve intercepts are \$100 and 50, and compute the equilibrium price and quantity in this market.

(b) If the government imposes a tax of \$10 per unit, draw the new equilibrium and compute the new quantity traded and the amount of tax revenue generated.

(c) Is demand elastic or inelastic in this price range? [*Hint*: you should be able to answer this without calculations, by observing the figure you have constructed.]

Exercise 4.11 *Optional*: The supply of Henry's hamburgers is given by $P = 2 + 0.5Q$; demand is given by $Q = 20$.

(a) Illustrate and compute the market equilibrium, knowing that the supply curve has an intercept of \$2 and a slope of 0.5.

(b) A specific tax of \$3 per unit is subsequently imposed and that shifts the supply curve upwards and parallel by \$3, to become $P = 5 + 0.5Q$. Solve for the equilibrium price and quantity after the tax.

(c) Insert the post-tax supply curve along with the pre-tax supply curve, and determine who bears the burden of the tax.

Chapter 5

Welfare economics and externalities

In this chapter we will explore:

5.1 Equity and efficiency

In modern mixed economies, markets and governments together determine the output produced and also who benefits from that output. In this chapter we explore a very broad question that forms the core of welfare economics: Even if market forces drive efficiency, are they a good way to allocate scarce resources in view of the fact that they not only give rise to inequality and poverty, but also fail to capture the impacts of productive activity on non-market participants? Mining impacts the environment, traffic results in road fatalities, alcohol, tobacco and opioids cause premature deaths. These products all generate secondary impacts beyond their stated objective. We frequently call these *external* effects.

The analysis of markets in this larger sense involves not just economic efficiency; public policy additionally has a normative content because policies can impact the various participants in different ways and to different degrees. **Welfare economics**, therefore, deals with both normative and positive issues.

> **Welfare economics** assesses how well the economy allocates its scarce resources in accordance with the goals of efficiency and equity.

Political parties on the left and right disagree on how well a market economy works. Canada's New Democratic Party emphasizes the market's failings and the need for government intervention, while the Progressive Conservative Party believes, broadly, that the market fosters choice, incentives, and efficiency. What lies behind this disagreement? The two principal factors are **efficiency** and **equity**. Efficiency addresses the question of how well the economy's resources are used and allocated. In contrast, equity deals with how society's goods and rewards are, and should be, distributed among its different members, and how the associated costs should be apportioned.

> **Equity** deals with how society's goods and rewards are, and should be, distributed among its different members, and how the associated costs should be apportioned.
>
> **Efficiency** addresses the question of how well the economy's resources are used and allocated.

Equity is also concerned with how different generations share an economy's productive capabilities: More investment today makes for a more productive economy tomorrow, but more greenhouse gases today will reduce environmental quality tomorrow. These are inter-generational questions.

Climate change caused by global warming forms one of the biggest challenges for humankind at the present time. As we shall see in this chapter, economics has much to say about appropriate policies to combat warming. Whether pollution-abatement policies should be implemented today or down the road involves considerations of equity between generations. Our first task is to develop an analytical tool which will prove vital in assessing and computing welfare benefits and costs – economic surplus.

5.2 Consumer and producer surplus

An understanding of economic efficiency is greatly facilitated as a result of understanding two related measures: Consumer surplus and producer surplus. Consumer surplus relates to the demand side of the market, producer surplus to the supply side. Producer surplus is also termed supplier surplus. These measures can be understood with the help of a standard example, the market for city apartments.

The market for apartments

Table 5.1 and Figure 5.1 describe the hypothetical data. We imagine first a series of city-based students who are in the market for a standardized downtown apartment. These individuals are not identical; they value the apartment differently. For example, Alex enjoys comfort and therefore places a higher value on a unit than Brian. Brian, in turn, values it more highly than Cathy or Don. Evan and Frank would prefer to spend their money on entertainment, and so on. These valuations are represented in the middle column of the demand panel in Table 5.1, and also in Figure 5.1 with the highest valuations closest to the origin. The valuations reflect the willingness to pay of each consumer.

Table 5.1: Consumer and supplier surpluses

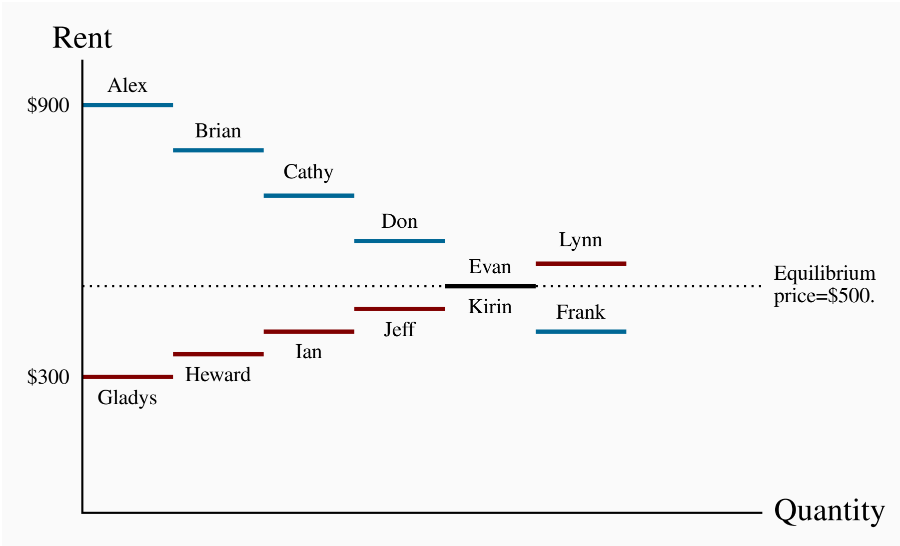
Demand		
Individual	**Demand valuation**	**Surplus**
Alex	900	400
Brian	800	300
Cathy	700	200
Don	600	100
Evan	500	0
Frank	400	0

Supply		
Individual	**Reservation value**	**Surplus**
Gladys	300	200
Heward	350	150
Ian	400	100
Jeff	450	50
Kirin	500	0
Lynn	550	0

On the supply side we imagine the market as being made up of different individuals or owners, who are willing to put their apartments on the market for different prices. Gladys will accept less rent than Heward, who in turn will accept less than Ian. The minimum prices that the suppliers are willing to accept are called *reservation* prices or values, and these are given in the lower part of Table 5.1. Unless the market price is greater than their reservation price, suppliers will hold back.

By definition, as stated in Chapter 3, the demand curve is made up of the valuations placed on the good by the various demanders. Likewise, the reservation values of the suppliers form the supply curve. If Alex is willing to pay \$900, then that is his demand price; if Heward is willing to put his apartment on the market for \$350, he is by definition willing to supply it for that price. Figure 5.1 therefore describes the demand and supply curves in this market. The steps reflect the willingness to pay of the buyers and the reservation valuations or prices of the suppliers.

Figure 5.1: The apartment market



Demanders and suppliers are ranked in order of the value they place on an apartment. The market equilibrium is where the marginal demand value of Evan equals the marginal supply value of Kirin at $500. Five apartments are rented in equilibrium.

In this example, the equilibrium price for apartments will be $500. Let us see why. At that price the value placed on the marginal unit supplied by Kirin equals Evan's willingness to pay. Five apartments will be rented. A sixth apartment will not be rented because Lynn will let her apartment only if the price reaches $550. But the sixth potential demander is willing to pay only $400. Note that, as usual, there is just a single price in the market. Each renter pays $500, and therefore each supplier also receives $500.

The consumer and supplier surpluses can now be computed. Note that, while Don is willing to pay $600, he actually pays $500. His consumer surplus is therefore $100. In Figure 5.1, we can see that each **consumer's surplus** is the distance between the market price and the individual's valuation. These values are given in the final column of the top half of Table 5.1.

> **Consumer surplus** is the excess of consumer willingness to pay over the market price.

Using the same reasoning, we can compute each **supplier's surplus**, which is the excess of the amount obtained for the rented apartment over the reservation price. For example, Heward obtains a surplus on the supply side of $150, while Jeff gets $50. Heward is willing to put his apartment on the market for $350, but gets the equilibrium price/rent of $500 for it. Hence his surplus is $150.

> **Supplier or producer surplus** is the excess of market price over the reservation price of the supplier.

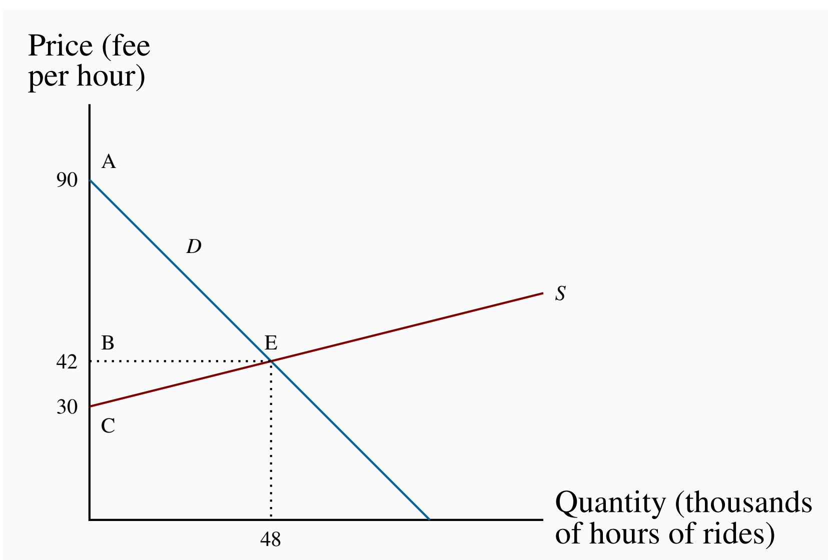
It should now be clear why these measures are called surpluses. *The suppliers and demanders are all willing to participate in this market because they earn this surplus*. It is a measure of their gain from being involved in the trading. The sum of each participant's surplus in the final column of Table 5.1 defines the total surplus in the market. Hence, on the demand side a total surplus arises of $1,000 and on the supply side a value of $500.

The taxi market

We do not normally think of demand and supply functions in terms of the steps illustrated in Figure 5.1. Usually there are so many participants in the market that the differences in reservation prices on the supply side and willingness to pay on the demand side are exceedingly small, and so the demand and supply curves are drawn as continuous lines. So our second example reflects this, and comes from the market for taxi rides. We might think of this as an *Uber*- or *Lyft*-type taxi operation.

Let us suppose that the demand and supply curves for taxi rides in a given city are given by the functions in Figure 5.2.

Figure 5.2: The taxi market



Consumer surplus is the area ABE, supplier surplus is the area BCE.

The demand curve represents the willingness to pay on the part of riders. The supply curve represents the willingness to supply on the part of drivers. The price per hour of rides defines the vertical axis; hours of rides (in thousands) are measured on the horizontal axis. The demand intercept of $90 says that the person who values the ride most highly is willing to pay $90 per hour.

The downward slope of the demand curve states that other buyers are willing to pay less. On the supply side no driver is willing to supply his time and vehicle unless he obtains at least \$30 per hour. To induce additional suppliers a higher price must be paid, and this is represented by the upward sloping supply curve.

The intersection occurs at a price of \$42 per hour and the equilibrium number of ride-hours supplied is 48 thousand[1]. Computing the surpluses is very straightforward. By definition the consumer surplus is the excess of the willingness to pay by each buyer above the uniform price. Buyers who value the ride most highly obtain the biggest surplus – the highest valuation rider gets a surplus of \$48 per hour – the difference between his willingness to pay of \$90 and the actual price of \$42. Each successive rider gets a slightly lower surplus until the final rider, who obtains zero. She pays \$42 and values the ride hours at \$42 also. On the supply side, the drivers who are willing to supply rides at the lowest reservation price (\$30 and above) obtain the biggest surplus. The 'marginal' supplier gets no surplus, because the price equals her reservation price.

From this discussion it follows that the consumer surplus is given by the area ABE and the supplier surplus by the area CBE. These are two triangular areas, and measured as half of the base by the perpendicular height. Therefore, in thousands of units:

$$\begin{aligned}\text{Consumer Surplus} &= \text{(demand value - price) = area ABE}\\ &= (1/2)\times 48\times \$48 = \$1,152\\ \text{Producer Surplus} &= \text{(price - reservation supply value) = area BEC}\\ &= (1/2)\times 48\times \$12 = \$288\end{aligned}$$

The total surplus that arises in the market is the sum of producer and consumer surpluses, and since the units are in thousands of hours the total surplus here is $(\$1,152+\$288)\times 1,000 = \$1,440,000$.

5.3 Efficient market outcomes

The definition and measurement of the surplus is straightforward provided the supply and demand functions are known. An important characteristic of the marketplace is that in certain circumstances it produces what we call an efficient outcome, or an **efficient market**. Such an outcome yields the highest possible sum of surpluses.

> An **efficient market** maximizes the sum of producer and consumer surpluses.

To see that this outcome achieves the goal of maximizing the total surplus, consider what would happen if the quantity $Q = 48$ in the taxi example were not supplied. Suppose that the city's taxi

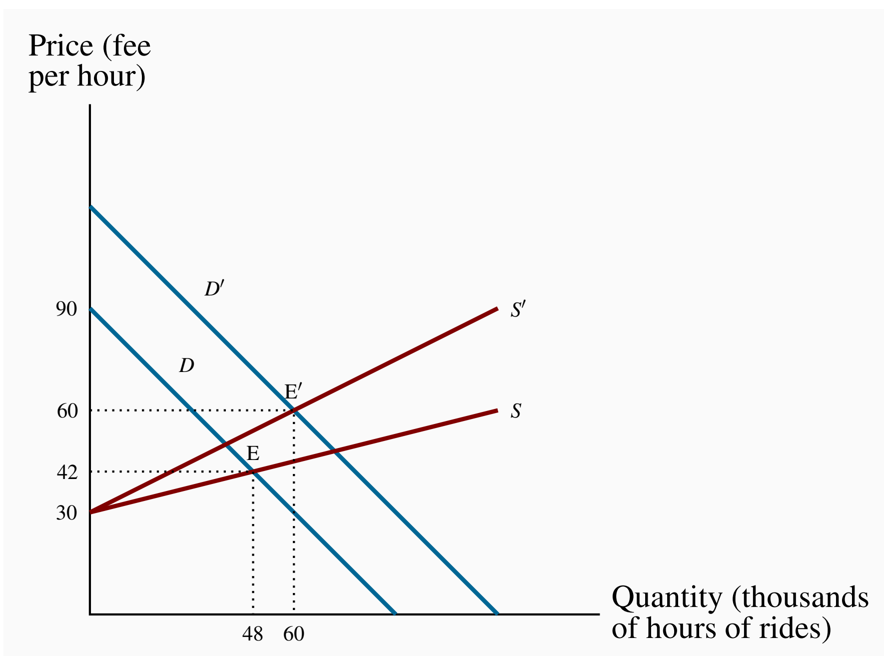
[1] The demand and supply functions behind these curves are $P = 90 - 1Q$ and $P = 30 + (1/4)Q$. Equating supply and demand yields the solutions in the text.

czar decreed that 50 units should be supplied, and the czar forced additional drivers on the road. If 2 additional units are to be traded in the market, consider the value of this at the margin. Suppliers value the supply more highly than the buyers are willing to pay. So on these additional 2 units negative surplus would accrue, thus reducing the total.

A second characteristic of the market equilibrium is that potential buyers who would like a cheaper ride and drivers who would like a higher hourly payment do not participate in the market. On the demand side those individuals who are unwilling to pay $42/hour can take public transit, and on the supply side the those drivers who are unwilling to supply at $42/hour can allocate their time to alternative activities. Obviously, only those who participate in the market benefit from a surplus.

One final characteristic of surplus measurement should be emphasized. That is, the surplus number is not unique, it depends upon the economic environment. We can illustrate this easily using the taxi example. A well recognized feature of *Uber* taxi rides is that the price varies with road and weather conditions. Poor weather conditions mean that there is an increased demand, and poor road or weather conditions mean that drivers are less willing to supply their services – their reservation payment increases. This situation is illustrated in Figure 5.3.

Figure 5.3: The taxi market



The curves represented by D' and S' represent the curves for bad weather: Taxi rides are more highly valued on the demand side, and drivers must be paid more to supply in less favourable work conditions.

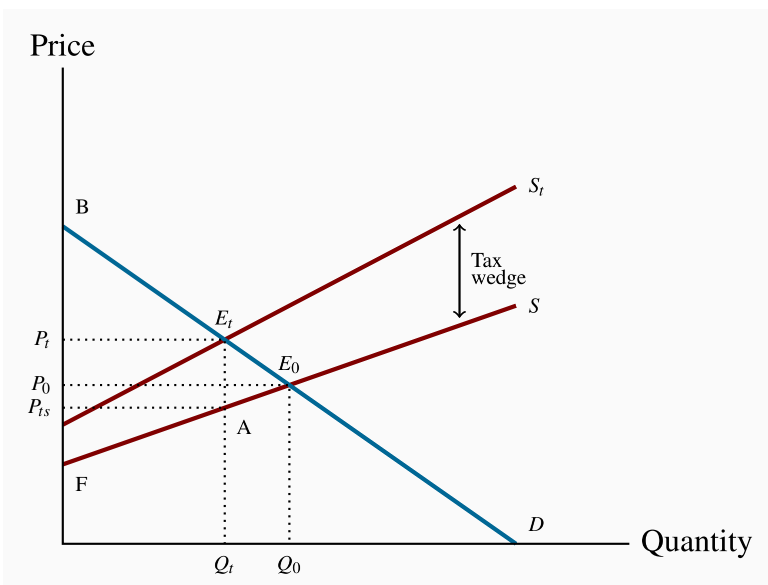
The demand curve has shifted upwards and the supply curve has also changed in such a way that any quantity will now be supplied at a higher price. The new equilibrium is given by E' rather than E.[2] There is a new equilibrium price-quantity combination that *is efficient in the new market conditions*. This illustrates that there is no such thing as a unique unchanging efficient outcome. When economic factors that influence the buyers' valuations (demand) or the suppliers' reservation prices (supply) change, then the efficient market outcome must be recomputed.

5.4 Taxation, surplus and efficiency

Despite enormous public interest in taxation and its impact on the economy, it is one of the least understood areas of public policy. In this section we will show how an understanding of two fundamental tools of analysis – elasticities and economic surplus – provides powerful insights into the field of taxation.

We begin with the simplest of cases: The federal government's goods and services tax (GST) or the provincial governments' sales taxes (PST). These taxes combined vary by province, but we suppose that a typical rate is 13 percent. In some provinces these two taxes are harmonized. Note that this is a *percentage*, or *ad valorem*, tax, not a *specific* tax of so many dollars per unit traded.

Figure 5.4: The efficiency cost of taxation



The tax shifts S to S_t and reduces the quantity traded from Q_0 to Q_t. At Q_t the demand value placed on an additional unit exceeds the supply valuation by E_tA. Since the tax keeps output at this lower level, the economy cannot take advantage of the additional potential surplus between Q_t and Q_0. Excess burden = deadweight loss = AE_tE_0.

[2] For example, if the demand curve shifts upwards, parallel, to become $P = 120 - 1Q$ and the supply curve changes in slope to $P = 30 + (1/2)Q$, the new equilibrium solution is $\{P = \$60, Q = 60\}$.

Figure 5.4 illustrates the supply and demand curves for some commodity. In the absence of taxes, the equilibrium E_0 is defined by the combination (P_0, Q_0).

A 13-percent tax is now imposed, and the new supply curve S_t lies 13 percent above the no-tax supply S. A **tax wedge** is therefore imposed between the price the consumer must pay and the price that the supplier receives. The new equilibrium is E_t, and the new market price is at P_t. The price received by the supplier is lower than that paid by the buyer by the amount of the tax wedge. The post-tax supply price is denoted by P_{ts}.

There are two *burdens* associated with this tax. The first is the **revenue burden**, the amount of tax revenue paid by the market participants and received by the government. On each of the Q_t units sold, the government receives the amount $(P_t - P_{ts})$. Therefore, tax revenue is the amount $P_tE_t\text{A}P_{ts}$. As illustrated in Chapter 4, the degree to which the market price P_t rises above the no-tax price P_0 depends on the supply and demand elasticities.

> A **tax wedge** is the difference between the consumer and producer prices.
>
> The **revenue burden** is the amount of tax revenue raised by a tax.

The second burden of the tax is called the *excess burden*. The concepts of consumer and producer surpluses help us comprehend this. The effect of the tax has been to reduce consumer surplus by $P_tE_tE_0P_0$. This is the reduction in the pre-tax surplus given by the triangle $P_0\text{B}E_0$. By the same reasoning, supplier surplus is reduced by the amount $P_0E_0\text{A}P_{ts}$; prior to the tax it was P_0E_0F. Consumers and suppliers have therefore seen a reduction in their well-being that is measured by these dollar amounts. Nonetheless, the government has additional revenues amounting to $P_tE_t\text{A}P_{ts}$, and this tax imposition therefore represents a *transfer* from the consumers and suppliers in the marketplace to the government. Ultimately, the citizens should benefit from this revenue when it is used by the government, and it is therefore not considered to be a net loss of surplus.

However, there remains a part of the surplus loss that is not transferred, the triangular area $E_tE_0\text{A}$. This component is called the **excess burden**, for the reason that it represents the component of the economic surplus that is not transferred to the government in the form of tax revenue. It is also called the **deadweight loss**, DWL.

> The **excess burden**, or **deadweight loss**, of a tax is the component of consumer and producer surpluses forming a net loss to the whole economy.

The intuition behind this concept is not difficult. At the output Q_t, the value placed by consumers on the last unit supplied is P_t $(= E_t)$, while the production cost of that last unit is P_{ts} (=A). But the potential surplus $(P_t - P_{ts})$ associated with producing an additional unit cannot be realized, because the tax dictates that the production equilibrium is at Q_t rather than any higher output. Thus, if output could be increased from Q_t to Q_0, a surplus of value over cost would be realized on every additional unit equal to the vertical distance between the demand and supply functions D and S. Therefore, the loss associated with the tax is the area E_tE_0A.

In public policy debates, this excess burden is rarely discussed. The reason is that notions of consumer and producer surpluses are not well understood by non-economists, despite the fact that the value of lost surpluses is frequently large. Numerous studies have estimated the excess burden associated with raising an additional dollar from the tax system. They rarely find that the excess burden is less than 25 percent of total expenditure. This is a sobering finding. It tells us that if the government wished to implement a new program by raising additional tax revenue, the benefits of the new program should be 25 percent greater than the amount expended on it!

The impact of taxes and other influences that result in an inefficient use of the economy's resources are frequently called **distortions** because they necessarily lead the economy away from the efficient output. The magnitude of the excess burden is determined by the elasticities of supply and demand in the markets where taxes are levied. To see this, return to Figure 5.4, and suppose that the demand curve through E_0 were more elastic (with the same supply curve, for simplicity). The post-tax equilibrium E_t would now yield a lower Q_t value and a price between P_t and P_0. The resulting tax revenue raised and the magnitude of the excess burden would differ because of the new elasticity.

> A **distortion** in resource allocation means that production is not at an efficient output.

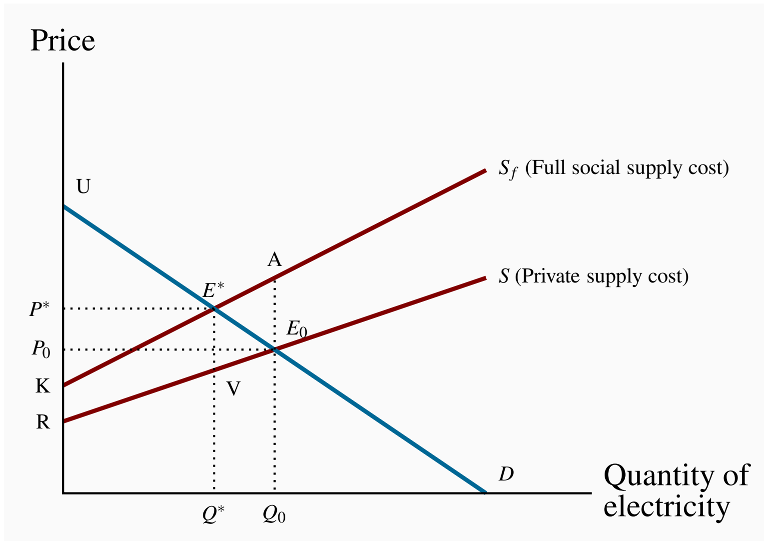
5.5 Market failures – externalities

The consumer and producer surplus concepts we have developed are extremely powerful tools of analysis, but the world is not always quite as straightforward as simple models indicate. For example, many suppliers generate pollutants that adversely affect the health of the population, or damage the environment, or both. The term **externality** is used to denote such impacts. Externalities impact individuals who are not participants in the market in question, and the effects of the externalities may not be captured in the market price. For example, electricity-generating plants that use coal reduce air quality, which, in turn, adversely impacts individuals who suffer from asthma or other lung ailments. While this is an example of a negative externality, externalities can also be positive.

> An **externality** is a benefit or cost falling on people other than those involved in the activity's market. It can create a difference between private costs or values and social costs or values.

We will now show why markets characterized by externalities are not efficient, and also show how these externalities might be corrected or reduced. The essence of an externality is that it creates a divergence between private costs/benefits and social costs/benefits. If a steel producer pollutes the air, and the steel buyer pays only the costs incurred by the producer, then the buyer is not paying the full "social" cost of the product. The problem is illustrated in Figure 5.5.

Figure 5.5: Negative externalities and inefficiency



A negative externality is associated with this good. S reflects private costs, whereas S_f reflects the full social cost. The socially optimal output is Q^*, not the market outcome Q_0. Beyond Q^* the real cost exceeds the demand value; therefore Q_0 is not an efficient output. A tax that increases P to P^* and reduces output is one solution to the externality.

Negative externalities

In Figure 5.5, the supply curve S represents the cost to the supplier, whereas S_f (the *full* cost) reflects, in addition, the cost of bad air to the population. Of course, we are assuming that this external cost is ascertainable, in order to be able to characterize S_f accurately. Note also that this illustration assumes that, as power output increases, the external cost *per unit* rises, because the difference between the two supply curves increases with output. This implies that low levels of pollution do less damage per unit: Perhaps the population has a natural tolerance for low levels, but higher levels cannot be tolerated easily and so the cost per unit is greater.

Despite the externality, *an efficient level of production can still be defined*. It is given by Q^*, not Q_0. To see why, consider the impact of reducing output by one unit from Q_0. At Q_0 the willingness of buyers to pay for the marginal unit supplied is E_0. The (private) supply cost is also E_0. But from a societal standpoint there is a pollution/health cost of AE_0 associated with that unit of production. The full cost, as represented by S_f, exceeds the buyer's valuation. Accordingly, if the last unit of output produced is cut, society gains by the amount AE_0, because the cut in output reduces the excess of true cost over value.

Applying this logic to each unit of output between Q_0 and Q^*, it is evident that society can increase its well-being by the dollar amount equal to the area E^*AE_0, as a result of reducing production.

Next, consider the consequences of reducing output further from Q^*. Note that some pollution is being created here, and environmentalists frequently advocate that pollution should be reduced to zero. However, an efficient outcome may not involve a zero level of pollution! If the production of power were reduced below Q^*, the loss in value to buyers, as a result of not being able to purchase the good, would exceed the full cost of its production.

If the government decreed that, instead of producing Q^*, no pollution would be tolerated, then society would forgo the possibility of earning the total real surplus equal to the area UE^*K. Economists do not advocate such a zero-pollution policy; rather, we advocate a policy that permits a "tolerable" pollution level – one that still results in net benefits to society. In this particular example, the total cost of the tolerated pollution equals the area between the private and full supply functions, KE^*VR.

As a matter of policy, how is this market influenced to produce the amount Q^* rather than Q_0? One option would be for the government to intervene directly with production quotas for each firm. An alternative would be to impose a **corrective tax** on the good whose production causes the externality: With an appropriate increase in the price, consumers will demand a reduced quantity. In Figure 5.5 a tax equal to the dollar value VE^* would shift the supply curve upward by that amount and result in the quantity Q^* being traded.

> A **corrective tax** seeks to direct the market towards a more efficient output.

We are now venturing into the field of environmental policy, where a corrective tax is usually called a carbon tax, and this is explored in the following section. The key conclusion of the foregoing analysis is that an efficient working of the market continues to have meaning in the presence of externalities. An efficient output level still maximizes economic surplus where surplus is correctly defined.

Positive externalities

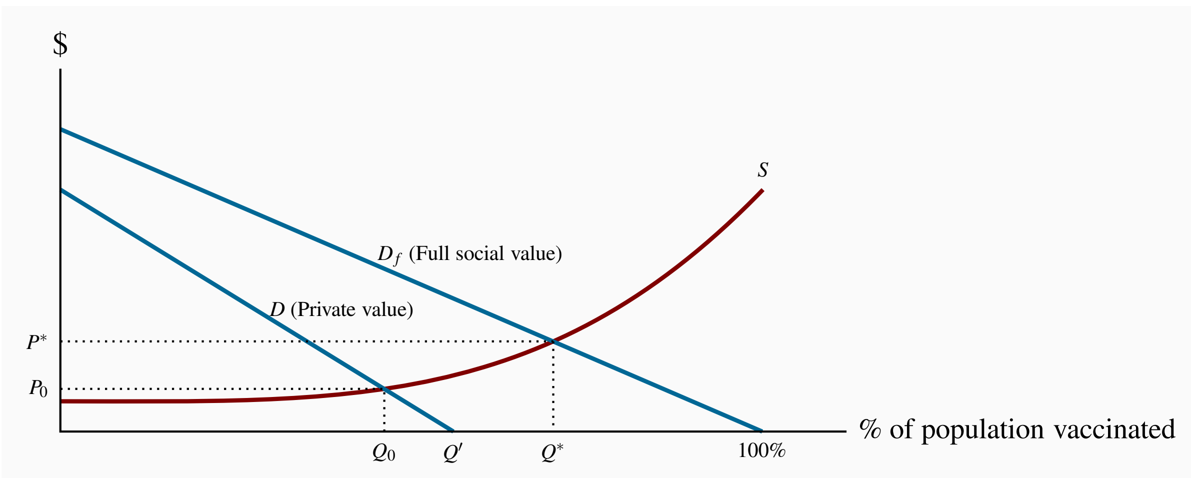
Externalities of the *positive* kind enable individuals or producers to get a type of 'free ride' on the efforts of others. Real world examples abound: When a large segment of the population is immunized against disease, the remaining individuals benefit on account of the reduced probability of transmission.

A less well recognized example is the benefit derived by many producers world-wide from research and development (R&D) undertaken in advanced economies and in universities and research institutes. The result is that society at large, including the corporate sector, gain from this enhanced understanding of science, the environment, or social behaviours.

The free market may not cope any better with these positive externalities than it does with negative externalities, and government intervention may be beneficial. Furthermore, firms that invest heavily in research and development would not undertake such investment if competitors could have a complete free ride and appropriate the fruits. This is why *patent laws* exist, as we shall see later in discussing Canada's competition policy. These laws prevent competitors from copying the product development of firms that invest in R&D. If such protection were not in place, firms would not allocate sufficient resources to R&D, which is a real engine of economic growth. In essence, the economy's research-directed resources would not be appropriately rewarded, and thus too little research would take place.

While patent protection is one form of corrective action, subsidies are another. We illustrated above that an appropriately formulated tax on a good that creates negative externalities can reduce demand for that good, and thereby reduce pollution. A subsidy can be thought of as a negative tax, and can stimulate the supply of goods and services that have positive externalities. Consider the example in Figure 5.6.

Figure 5.6: Positive externalities – the market for flu shots



The value to society of vaccinations exceeds the value to individuals: The greater the number of individuals vaccinated, the lower is the probability of others contracting the virus. D_f reflects this additional value. Consequently, the social optimum is Q^* which exceeds Q_0.

Individuals have a demand for flu shots given by D. This reflects their private valuation – their personal willingness to pay. But the social value of flu shots is greater. When more individuals are vaccinated, the probability that others will be infected falls. Additionally, with higher rates of immunization, the health system will incur fewer costs in treating the infected. Therefore, the value to society of any quantity of flu shots is greater than the sum of the values that individuals place on them.

D_f reflects the full social value of any quantity of flu shots. In this instance the quantity axis measures the percentage of the population vaccinated, which has a maximum of 100%. If S is the supply curve, the socially optimal, efficient, market outcome is Q^*. The steeply upward-sloping section of S denotes that it may be very costly to vaccinate every last person – particularly those living in outlying communities. How can we influence the market to move from Q_0 towards Q^*? One solution is a subsidy that would reduce the price to zero. In this case that gets us almost to the optimum, because the percentage of the population now choosing to be vaccinated is given by Q'. The zero price essentially makes the supply curve, as perceived by the population, to be running along the horizontal axis.

Note the social value of the improvement in moving from Q_0 to Q'; the social value exceeds the social cost. But even at Q' further gains are available because at Q' the social value of additional vaccinations is greater than the social cost. Overall, at the point Q^* the social value is given by the area under the demand curve, and the social cost by the area under the supply curve.

5.6 Other market failures

There are other ways in which markets can fail to reflect accurately the social value or social cost of economic activity. Profit-seeking monopolies, which restrict output in order to increase profits, represent inefficient markets, and we will see why in the chapter on monopoly. Or the market may not deal very well with what are called public goods. These are goods, like radio and television service, national defence, or health information: With such goods and services many individuals can be supplied with the same good at the same total cost as one individual. We will address this problem in our chapter on government. And, of course, there are international externalities that cannot be corrected by national governments because the interests of adjoining states may differ: One economy may wish to see cheap coal-based electricity being supplied to its consumers, even if this means acid rain or reduced air quality in a neighbouring state. Markets may fail to supply an "efficient" amount of a good or service in all of these situations. Global warming is perhaps the best, and most extreme, example of international externalities and market failure.

5.7 Environmental policy and climate change

Greenhouse gases

The greatest externality challenge in the modern world is to control our emissions of greenhouse gases.The emission of **greenhouse gases** (GHGs) is associated with a wide variety of economic activities such as coal-based power generation, oil-burning motors, wood-burning stoves, ruminant animals, *etc*. The most common GHG is carbon dioxide, methane is another. The gases, upon emission, circulate in the earth's atmosphere and, following an excessive build-up, prevent sufficient radiant heat from escaping. The result is a slow warming of the earth's surface and air temperatures. It is envisaged that such temperature increases will, in the long term, increase water temperatures and cause glacial melting, with the result that water levels worldwide will rise. In addition to the higher water levels, which the Intergovernmental Panel on Climate Change (IPCC) estimates will be between one foot and one metre by the end of the 21st century, oceans will become more acidic, weather patterns will change and weather events become more variable and severe. The changes will be latitude-specific and vary by economy and continent, and ultimately will impact the agricultural production abilities of certain economies.

> **Greenhouse gases** that accumulate excessively in the earth's atmosphere prevent heat from escaping and lead to **global warming**.

While most scientific findings and predictions are subject to a degree of uncertainty, there is little disagreement in the scientific community on the long-term impact of increasing GHGs in the atmosphere. There is some skepticism as to whether the generally higher temperatures experienced in recent decades are completely attributable to anthropogenic activity since the industrial revolution, or whether they also reflect a natural cycle in the earth's temperature. But scientists agree that a continuance of the recent rate of GHG emissions is leading to serious climatic problems.

The major economic environmental challenge facing the world economy is this: Historically, GHG emissions have been strongly correlated with economic growth. The very high rate of economic growth in many large-population economies such as China and India that will be necessary to raise hundreds of millions out of poverty means that that historical pattern needs to be broken – GHG accumulation must be "decoupled" from economic growth.

GHGs as a common property

A critical characteristic of GHGs is that they are what we call in economics a 'common property': Every citizen in the world 'owns' them, every citizen has equal access to them, and it matters little where these GHGs originate. Consequently, if economy A reduces its GHG emissions, economy B may simply increase its emissions rather than incur the cost of reducing them. Hence, economy A's behaviour goes unrewarded. This is the crux of international agreements – or disagreements. Since GHGs are a common property, in order for A to have the incentive to reduce emissions, it needs to know that B will act correspondingly.

From the Kyoto Protocol to the Paris Accord

The world's first major response to climate concerns came in the form of the United Nations–sponsored Earth Summit in Rio de Janeiro in 1992. This was followed by the signing of the Kyoto Protocol in 1997, in which a group of countries committed themselves to reducing their GHG emissions relative to their 1990 emissions levels by the year 2012. Canada's Parliament subsequently ratified the Kyoto Protocol, and thereby agreed to meet Canada's target of a 6 percent reduction in GHGs relative to the amount emitted in 1990.

On a per-capita basis, Canada is one of the world's largest contributors to global warming, even though Canada's percentage of the total is just 2 percent. Many of the world's major economies refrained from signing the Protocol—most notably China, the United States, and India. Canada's emissions in 1990 amounted to approximately 600 giga tonnes (Gt) of carbon dioxide; but by the time we ratified the treaty in 2002, emissions were 25% above that level. Hence the signing was somewhat meaningless, in that Canada had virtually a zero possibility of attaining its target.

The target date of 2012 has come and gone and subsequent conferences in Copenhagen and Rio failed to yield an international agreement. But in Paris, December 2015, 195 economies committed to reduce their GHG emissions by specific amounts. Canada was a party to that agreement. Target reductions varied by country. Canada committed itself to reduce GHG emissions by 30% by the year 2030 relative to 2005 emissions levels. To this end the Liberal government of Prime Minister Justin Trudeau announced in late 2016 that if individual Canadian provinces failed to implement a carbon tax, or equivalent, the federal government would impose one unilaterally. The program involves a carbon tax of $10 per tonne in 2018, that increases by $10 per annum until it attains a value of $50 in 2022. Some provinces already have GHG limitation systems in place (cap and trade systems - developed below), and these provinces would not be subject to the federal carbon tax provided the province-level limitation is equivalent to the federal carbon tax.

Canada's GHG emissions

An excellent summary source of data on Canada's emissions and performance during the period 1990-2018 is available on Environment Canada's web site. See:

https://www.canada.ca/en/environment-climate-change/services/climate-change/greenhouse-gas-emissions/sources-sinks-executive-summary-2020.html#toc3

Canada, like many economies, has become more efficient in its use of energy (the main source of GHGs) in recent decades—its *use of energy per unit of total output* has declined steadily. Canada emitted 0.44 mega tonnes of CO_2 equivalent per billion dollars of GDP in 2005, and 0.36 mega tonnes in 2017. On a *per capita* basis Canada's emissions amounted to 22.9 tonnes in 2005, and dropped to 19.5 by 2017. This modest improvement in efficiency means that Canada's GDP is now *less energy intensive*. The critical challenge is to produce more output while using not just less energy per unit of output, but to use less energy in total.

While Canada's energy intensity (GHGs per unit of output) has dropped, overall emissions have increased by almost 20% since 1990. Furthermore, while developed economies have increased their efficiency, it is the *world's* efficiency that is ultimately critical. By outsourcing much of its manufacturing sector to China, Canada and the West have offloaded some of their most GHG-intensive activities. But GHGs are a common property resource.

Canada's GHG emissions also have a regional aspect: The *production* of oil and gas, which has created considerable wealth for all Canadians, is both energy intensive and concentrated in a limited number of provinces (Alberta, Saskatchewan and more recently Newfoundland and Labrador).

GHG measurement

GHG atmospheric concentrations are measured in parts per million (ppm). Current levels in the atmosphere are slightly above 400 ppm, and continued growth in concentration will lead to serious economic and social disruption. In the immediate pre-industrial revolution era concentrations were in the 280 ppm range. Hence, our world seems to be headed towards a doubling of GHG concentrations in the coming decades.

GHGs are augmented by the annual additions to the stock already in the atmosphere, and at the same time they decay—though very slowly. GHG-reduction strategies that propose an immediate reduction in emissions are more costly than those aimed at a more gradual reduction. For example, a slower investment strategy would permit in-place production and transportation equipment to reach the end of its economic life rather than be scrapped and replaced 'prematurely'. Policies that focus upon longer-term replacement are therefore less costly in this specific sense.

While not all economists and policy makers agree on the time scale for attacking the problem, the longer that GHG reduction is postponed, the greater the efforts will have to be in the long term—because GHGs will build up more rapidly in the near term.

A critical question in controlling GHG emissions relates to the cost of their control: How much of annual growth might need to be sacrificed in order to get emissions onto a sustainable path? Again estimates vary. The Stern Review (2006) proposed that, with an increase in technological capabilities, a strategy that focuses on the relative near-term implementation of GHG reduction measures might cost "only" a few percentage points of the value of world output. If correct, this is a low price to pay for risk avoidance in the longer term.

Nonetheless, such a reduction will require particular economic policies, and specific sectors will be impacted more than others.

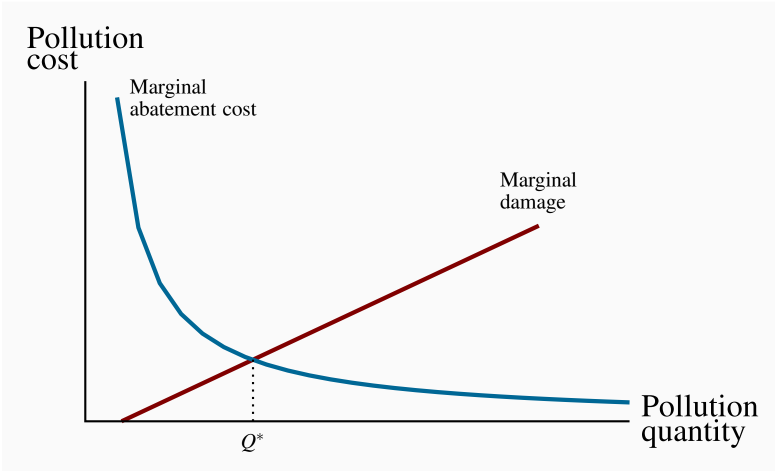
Economic policies for climate change

There are three main ways in which polluters can be controlled. One involves issuing direct controls; the other two involve incentives—in the form of pollution taxes, or on tradable "permits" to pollute.

To see how these different policies operate, consider first Figure 5.7. It is a standard diagram in environmental economics, and is somewhat similar to our supply and demand curves. On the horizontal axis is measured the quantity of environmental damage or pollution, and on the vertical axis its dollar value or cost. The upward-sloping damage curve represents the cost to society of each additional unit of pollution or gas, and it is therefore called a **marginal damage curve**. It is positively sloped to reflect the reality that, at low levels of emissions, the damage of one more unit is less than at higher levels. In terms of our earlier discussion, this means that an increase in GHGs of 10 ppm when concentrations are at 300 ppm may be less damaging than a corresponding increase when concentrations are at 500 ppm.

> The **marginal damage curve** reflects the cost to society of an additional unit of pollution.

Figure 5.7: The optimal quantity of pollution



Q^* represents the optimal amount of pollution. More than this would involve additional social costs because damages exceed abatement costs. Conversely, less than Q^* would require an abatement cost that exceeds the reduction in damage.

The second curve is the abatement curve. It reflects the cost of reducing emissions by one unit, and is therefore called a **marginal abatement curve**. This curve has a negative slope indicating that, as we reduce the total quantity of pollution produced (moving towards the origin on the horizontal axis), the cost of further unit reductions rises. This shape corresponds to reality. For example, halving the emissions of pollutants and gases from automobiles may be achieved by adding a catalytic converter and reducing the amount of lead in gasoline. But reducing those emissions all the way to zero requires the development of major new technologies such as electric cars—an enormously more costly undertaking.

> The **marginal abatement curve** reflects the cost to society of reducing the quantity of pollution by one unit.

If producers are unconstrained in the amount of pollution they produce, they will produce more than what we will show is the optimal amount – corresponding to Q^*. This amount is optimal in the sense that at levels greater than Q^* the damage exceeds the cost of reducing the emissions. However, reducing emissions below Q^* would mean incurring a cost per unit reduction that exceeds the benefit of that reduction. Another way of illustrating this is to observe that at a level of pollution above Q^* the cost of reducing it is less than the damage it inflicts, and therefore a net gain accrues to society as a result of the reduction. But to reduce pollution below Q^* would involve an abatement cost greater than the reduction in pollution damage and therefore no net gain to society. This constitutes a first rule in optimal pollution policy.

An optimal quantity of pollution occurs when the marginal cost of abatement equals the marginal damage.

A second guiding principle emerges by considering a situation in which some firms are relatively 'clean' and others are 'dirty'. More specifically, a clean firm A may have already invested in new equipment that uses less energy per unit of output produced, or emits fewer pollutants per unit of output. In contrast, the dirty firm B uses older dirtier technology. Suppose furthermore that these two firms form a particular sector of the economy and that the government sets a limit on total pollution from this sector, and that this limit is less than what the two firms are currently producing. What is the least costly method to meet the target?

The intuitive answer to this question goes as follows: In order to reduce pollution at least cost to the sector, calculate what it would cost each firm to reduce pollution from its present level. Then implement a system so that the firm with the least cost of reduction is the first to act. In this case the 'dirty' firm will likely have a lower cost of abatement since it has not yet upgraded its physical plant. This leads to a second rule in pollution policy:

With many polluters, the least cost policy to society requires producers with the lowest abatement costs to act first.

This principle implies that policies which impose the same emission limits on firms may not be the least costly manner of achieving a target level of pollution. Let us now consider the use of **tradable permits** and **corrective/carbon taxes** as policy instruments. These are market-based systems aimed at reducing GHGs.

> **Tradable permits** and **corrective/carbon taxes** are market-based systems aimed at reducing GHGs.

Incentive mechanism I: Tradable permits

A system of tradable permits is frequently called a 'cap and trade' system, because it limits or caps the total permissible emissions, while at the same time allows a market to develop in permits. For illustrative purposes, consider the hypothetical two-firm sector we developed above, composed of firms A and B. Firm A has invested in clean technology, firm B has not. Thus it is less costly for B to reduce emissions than A if further reductions are required. Next suppose that each firm is allocated by the government a specific number of 'GHG emission permits'; and that the total of such permits is less than the amount of emissions at present, and that each firm is emitting more than its permits allow. How can these firms achieve the target set for this sector of the economy?

The answer is that they should be able to engage in mutually beneficial trade: If firm B has a lower cost of reducing emissions than A, then it may be in A's interest to pay B to reduce B's emissions heavily. Imagine that each firm is emitting 60 units of GHG, but they have permits to emit only 50 units each. And furthermore suppose it costs B $20 to reduce GHGs by one unit, whereas it costs A $30 to do this. In this situation A could pay B $25 for several permits and this would benefit both firms. B can reduce GHGs at a cost of $20 and is being paid $25 to do this. In turn A would incur a cost of $30 per unit to reduce his GHGs but he can buy permits from B for just $25 and avoid the $30 cost. Both firms gain, and the total cost to the economy is lower than if each firm had to reduce by the same amount.

The benefit of the cap 'n trade system is that it enables the marketplace to reduce GHGs at least cost.

The largest system of tradable permits currently operates in the European Union: The *EU Emissions Trading System*. It covers more than 10,000 large energy-using installations. Trading began in 2005. In North America a number of Western states and several Canadian provinces are joined, either as participants or observers, in the *Western Climate Initiative*, which is committed to reduce GHGs by means of tradable emissions permits. The longer-term goal of these systems is for the government to issue progressively fewer permits each year, and to include an ever larger share of GHG-emitting enterprises with the passage of time.

Policy in practice – international

In an ideal world, permits would be traded internationally, and such a system might be of benefit to developing economies: If the cost of reducing pollution is relatively low in developing economies because they have few controls in place, then developed economies, for whom the cost of GA reduction is high, could induce firms in the developing world to undertake cost reductions. Such a trade would be mutually beneficial. For example, imagine in the above example that B is located in the developing world and A in the developed world. Both would obviously gain from such an arrangement, and because GHGs are a common property, the source of GHGs from a damage standpoint is immaterial.

Incentive mechanism II: Taxes

Corrective taxes are frequently called *Pigovian* taxes, after the economist Arthur Pigou. He advocated taxing activities that cause negative externalities. These taxes have been examined above in Section 5.4. Corrective taxes of this type can be implemented as part of a tax package *reform*. For example, taxpayers are frequently reluctant to see governments take 'yet more' of their money, in the form of new taxes. Such concerns can be addressed by reducing taxes in other sectors of the economy, in such a way that the package of tax changes maintains a 'revenue neutral' impact.

Revenues from taxes and permits

Taxes and tradable permits differ in that taxes generate revenue for the government from polluting producers, whereas permits may not generate revenue, or may generate less revenue. If the government simply *allocates* permits initially to all polluters, free of charge, and allows a market to develop, such a process generates no revenue to the government. While economists may advocate an *auction* of permits in the start-up phase of a tradable permits market, such a mechanism may run into political objections.

Setting taxes at the appropriate level requires knowledge of the cost and damage functions associated with GHGs. At the present time, economists and environmental scientists think that an appropriate price or tax on one tonne of GHG is in the $40 – $50 range. Such a tax would reduce emissions to a point where the longer-term impact of GHGs would not be so severe as otherwise.

British Columbia introduced a carbon tax of $10 per tonne of GHG on fuels in 2008, and has increased that price regularly. This tax was designed to be revenue neutral in order to make it more acceptable. This means that British Columbia reduced its income tax rates by an amount such that income tax payments would fall by an amount equal to the revenue captured by the carbon tax.

GHG policy at the federal level in Canada is embodied in the *Greenhouse Gas Pollution Pricing Act of 2018*. As detailed earlier, the Act imposes a yearly increasing levy on emissions. The system is intended to be revenue neutral, in that the revenues will be returned to households in the form of a 'paycheck' by the federal government. Large emitters of GHGs are permitted a specific threshold number of tonnes of emission each year without being penalized. Beyond that threshold the above rates apply.

Will this amount of carbon taxation hurt consumers, and will it enable Canada to reach its 2030 GHG goal? As a specific example: the gasoline-pricing rule of thumb is that each $10 in carbon taxation or pricing leads to an increase in the price of gasoline at the pump of about 2.5 cents. So a $50 levy per tonne means gas at the pump should rise by 12.5 cents per litre. The proceeds are returned to households.

As for the goal of reaching the 2030 target announced at Paris: Environment Canada estimates that the pricing scheme will reduce GHG emissions by about 60 tonnes per annum. But Canada's goal stated in Paris is to reduce emissions in 2030 by approximately four times this amount. Under the Paris Accoord, Canada stated that its 2030 goal would be to reduce emissions by 30% from their 2005 level of 725 MT, that is by an amount equal to approximately 220 tonnes.

Policy in practice – domestic large final emitters

Governments frequently focus upon quantities emitted by individual large firms, or *large final emitters (LFEs)*. In some economies, a relatively small number of producers are responsible for a disproportionate amount of an economy's total pollution, and limits are placed on those firms in the belief that significant economy-wide reductions can be achieved in this manner. One reason for concentrating on these LFEs is that the monitoring costs are relatively small compared to the costs associated with monitoring *all* firms in the economy. It must be kept in mind that pollution permits may be a legal requirement in some jurisdictions, but monitoring is still required, because firms could choose to risk polluting without owning a permit.

Conclusion

Welfare economics lies at the heart of public policy. Demand and supply curves can be interpreted as value curves and cost curves when there are no externalities involved. This is what enables us to define an efficient output of a product, and consequently an efficient use of the economy's resources. While efficiency is a central concept in economics, we must keep in mind that when the economic environment changes so too will the efficient use of resources, as we illustrated in

Section 5.7.

In this chapter we have focused on equity issues through the lens of GHG emissions. The build-up of GHGs in our atmosphere invokes the concept of *intergenerational equity*: The current generation is damaging the environment and the costs of that damage will be borne by subsequent generations. Hence it is inequitable in the intergenerational sense for us to leave a negative legacy to succeeding generations. Equity arises within generations also. For example, how much more in taxes should the rich pay relative to the non-rich? We will explore this type of equity in our chapter on government.

Key Terms

Welfare economics assesses how well the economy allocates its scarce resources in accordance with the goals of efficiency and equity.

Efficiency addresses the question of how well the economy's resources are used and allocated.

Equity deals with how society's goods and rewards are, and should be, distributed among its different members, and how the associated costs should be apportioned.

Consumer surplus is the excess of consumer willingness to pay over the market price.

Supplier or producer surplus is the excess of market price over the reservation price of the supplier.

Efficient market: maximizes the sum of producer and consumer surpluses.

Tax wedge is the difference between the consumer and producer prices.

Revenue burden is the amount of tax revenue raised by a tax.

Excess burden of a tax is the component of consumer and producer surpluses forming a net loss to the whole economy.

Deadweight loss of a tax is the component of consumer and producer surpluses forming a net loss to the whole economy.

Distortion in resource allocation means that production is not at an efficient output.

Externality is a benefit or cost falling on people other than those involved in the activity's market. It can create a difference between private costs or values and social costs or values.

Corrective tax seeks to direct the market towards a more efficient output.

Greenhouse gases that accumulate excessively in the earth's atmosphere prevent heat from escaping and lead to global warming.

Marginal damage curve reflects the cost to society of an additional unit of pollution.

Marginal abatement curve reflects the cost to society of reducing the quantity of pollution by one unit.

Tradable permits are a market-based system aimed at reducing GHGs.

Carbon taxes are a market-based system aimed at reducing GHGs.

Exercises for Chapter 5

Exercise 5.1 Four teenagers live on your street. Each is willing to shovel snow from one driveway each day. Their "willingness to shovel" valuations (supply) are: Jean, \$10; Kevin, \$9; Liam, \$7; Margaret, \$5. Several households are interested in having their driveways shoveled, and their willingness to pay values (demand) are: Jones, \$8; Kirpinsky, \$4; Lafleur, \$7.50; Murray, \$6.

(a) Draw the implied supply and demand curves as step functions.

(b) How many driveways will be shoveled in equilibrium?

(c) Compute the maximum possible sum for the consumer and supplier surpluses.

(d) If a new (wealthy) family arrives on the block, that is willing to pay \$12 to have their driveway cleared, recompute the answers to parts (a), (b), and (c).

Exercise 5.2 Consider a market where supply curve is horizontal at $P = 10$ and the demand curve has intercepts $\{\$34, 34\}$, and is defined by the relation $P = 34 - Q$.

(a) Illustrate the market geometrically.

(b) Impose a tax of \$2 per unit on the good so that the supply curve is now $P = 12$. Illustrate the new equilibrium quantity.

(c) Illustrate in your diagram the tax revenue generated.

(d) Illustrate the deadweight loss of the tax.

Exercise 5.3 Next, consider an example of DWL in the labour market. Suppose the demand for labour is given by the fixed gross wage $W = \$16$. The supply is given by $W = 0.8L$, indicating that the supply curve goes through the origin with a slope of 0.8.

(a) Illustrate the market geometrically.

(b) Calculate the supplier surplus, knowing that the equilibrium is $L = 20$.

(c) *Optional*: Suppose a wage tax is imposed that produces a net-of-tax wage equal to $W = \$12$. This can be seen as a downward shift in the demand curve. Illustrate the new quantity supplied and the new supplier's surplus.

Exercise 5.4 Governments are in the business of providing information to potential buyers. The first serious provision of information on the health consequences of tobacco use appeared in the United States Report of the Surgeon General in 1964.

(a) How would you represent this intervention in a supply and demand for tobacco diagram?

(b) Did this intervention "correct" the existing market demand?

Exercise 5.5 In deciding to drive a car in the rush hour, you think about the cost of gas and the time of the trip.

(a) Do you slow down other people by driving?

(b) Is this an externality, given that you yourself are suffering from slow traffic?

Exercise 5.6 Suppose that our local power station burns coal to generate electricity. The demand and supply functions for electricity are given by $P = 12 - 0.5Q$ and $P = 2 + 0.5Q$, respectively. The demand curve has intercepts $\{\$12, 24\}$ and the supply curve intercept is at \$2 with a slope of one half. However, for each unit of electricity generated, there is an externality. When we factor this into the supply side of the market, the real social cost is increased by \$1 per unit. That is, the supply curve shifts upwards by \$1, and now takes the form $P = 3 + 0.5Q$.

(a) Illustrate the free-market equilibrium.

(b) Illustrate the efficient (i.e. socially optimal) level of production.

Exercise 5.7 Your local dry cleaner, Bleached Brite, is willing to launder shirts at its cost of \$1.00 per shirt. The neighbourhood demand for this service is $P = 5 - 0.005Q$, knowing that the demand intercepts are $\{\$5, 1000\}$.

(a) Illustrate the market equilibrium.

(b) Suppose that, for each shirt, Bleached Brite emits chemicals into the local environment that cause \$0.25 damage per shirt. This means the full cost of each shirt is \$1.25. Illustrate graphically the socially optimal number of shirts to be cleaned.

(c) *Optional*: Calculate the socially optimal number of shirts to be cleaned.

Exercise 5.8 The supply curve for agricultural labour is given by $W = 6 + 0.1L$, where W is the wage (price per unit) and L the quantity traded. Employers are willing to pay a wage of \$12 to all workers who are willing to work at that wage; hence the demand curve is $W = 12$.

(a) Illustrate the market equilibrium, if you are told that the equilibrium occurs where $L = 60$.

(b) Compute the supplier surplus at this equilibrium.

Exercise 5.9 *Optional*: The market demand for vaccine XYZ is given by $P = 36 - Q$ and the supply conditions are $P = 20$; so \$20 represents the true cost of supplying a unit of vaccine. There is a positive externality associated with being vaccinated, and the real societal value is known and given by $P = 36 - (1/2)Q$. This new demand curve represents the true value to society of each vaccination. This is reflected in the private value demand curve rotating upward around the price intercept of \$36.

(a) Illustrate the private and social demand curves on a diagram, with intercept values calculated.

(b) What is the market solution to this supply and demand problem?

(c) What is the socially optimal number of vaccinations?

Part Three

Decision Making by Consumer and Producers

Individuals and producers make choices when they interact with one another. We begin by supposing that individuals act in a manner that is at least in their self-interest – they transact in markets because they derive satisfaction or utility from doing so. Individuals may also be altruistic. But in each case we suppose they plan their actions in a way that is consistent with a goal. Utility may be measurable or only comparable, and we derive the tools of choice optimization using these two alternative notions of utility in Chapter 6.

Chapter 7 furnishes a link between buyers and sellers, between savers and investors, between households and firms. We explore the workings of corporations, and illustrate how investment in many different firms can reduce risk for an individual investor.

Like individual buyers, sellers and producers also act with a plan to reach a goal. They are interested in making a profit from their enterprise and produce goods and services at the least cost that is consistent with available production technologies. Cost minimization and profit maximization in the short run and the long run are explored in Chapter 8.

Chapter 6

Individual choice

In this chapter we will explore:

6.1 Rationality

A critical behavioural assumption in economics is that agents operate in a way that is oriented towards achieving a goal. This can be interpreted to mean that individuals and firms maximize their personal well-being and/or their profits. These players may have other goals in addition: Philanthropy and the well-being of others are consistent with individual optimization.

If individuals are to achieve their goals then they must act in a manner that will get them to their objective; broadly, they must act in a rational manner. The theory of individual maximization that we will develop in this chapter is based on that premise or assumption. In assuming individuals are rational we need not assume that they have every piece of information available to them that might be relevant for a specific decision or choice. Nor need we assume that they have super computers in their brain when they evaluate alternative possible strategies.

What we do need to assume, however, is that individuals act in a manner that is consistent with obtaining a given objective. The modern theory of behavioural economics and behavioural psychology examines decision making in a wide range of circumstances and has uncovered many fascinating behaviours – some of which are developed in Application Box 6.1 below.

We indicated in Chapter 1 that as social scientists, we require a *reliable model* of behaviour, that is, *a way of describing the essentials of choice that is consistent with everyday observations on individual behaviour patterns*. In this chapter, our aim is to understand more fully the behavioural forces that drive the demand side of the economy.

Economists analyze individual decision making using two different, yet complementary, approaches – utility analysis and indifference analysis. We begin by portraying individuals as maximizing their *measurable utility* (sometimes called *cardinal utility*); then progress to indifference analysis, where a weaker assumption is made on the ability of individuals to measure their satisfaction. In this second instance we do not assume that individuals can measure their utility numerically, only that they can say if one collection of goods and services yields them greater satisfaction than another group. This ranking of choices corresponds to what is sometimes called *ordinal utility* – because individuals can *order* groups of goods and services in ascending order of satisfaction. In each case individuals are perceived as rational maximizers or optimizers: They allocate their income so as to choose the outcome that will make them as well off as possible.

The second approach to consumer behaviour is frequently omitted in introductory texts. It can be omitted here without interpreting the flow of ideas, although it does yield additional insights into consumer choice and government policy. As in preceding chapters, we begin the analysis with a motivating numerical example.

Application Box 6.1: Rationality and impulse

A number of informative and popular books on decision making have appeared recently. Their central theme is that our decision processes should not be viewed solely as a rational computer – operating in one single mode only, and unmoved by our emotions or history. Psychologists now know that our brains have at least two decision modes, and these are developed by economics Nobel Prize winner Daniel Kahneman in his book "Thinking, Fast and Slow". One part of our brain operates in a rational goal-oriented forward-looking manner (the 'slow' part), another is motivated by immediate gratification (the 'fast' part). Decisions that we observe in the world about us reflect these different mechanisms.

Richard Thaler, a Chicago economist and his law professor colleague Cass Sunstein, have developed a role for public policy in their book entitled "Nudge". They too argue that individuals do not inevitably operate in their own best long-term interests, and as a consequence individuals frequently require a *nudge* by government to make the long-term choice rather than the short-term choice. For example, when individuals begin a new job, they might be automatically enrolled in the company pension plan and be given the freedom to opt out, rather than not be enrolled and given the choice to opt in. Such policies are deemed to be 'soft paternalism'. They are paternalistic for the obvious reason – another organism is directing, but they are also soft in that they are not binding.

6.2 Choice with measurable utility

Neal loves to pump his way through the high-altitude powder at the Whistler ski and snowboard resort. His student-rate lift-ticket cost is $30 per visit. He also loves to frequent the jazz bars in downtown Vancouver, and each such visit costs him $20. With expensive passions, Neal must allocate his monthly entertainment budget carefully. He has evaluated how much satisfaction, measured in utils, he obtains from each snowboard outing and each jazz club visit. We assume that these utils are measurable, and use the term **cardinal utility** to denote this. These measurable utility values are listed in columns 2 and 3 of Table 6.1. They define the **total utility** he gets from various amounts of the two activities.

Table 6.1: Utils from snowboarding and jazz

1	2	3	4	5	6	7
Visit #	**Total snowboard utils**	**Total jazz utils**	**Marginal snowboard utils**	**Marginal jazz utils**	**Marginal snowboard utils per $**	**Marginal jazz utils per $**
1	72	52	72	52	2.4	2.6
2	132	94	60	42	2.0	2.1
3	182	128	50	34	1.67	1.7
4	224	156	42	28	1.4	1.4
5	260	180	36	24	1.2	1.2
6	292	201	32	21	1.07	1.05
7	321	220	29	19	0.97	0.95

Price of snowboard visit=$30. Price of jazz club visit=$20.

> **Cardinal utility** is a measurable concept of satisfaction.
>
> **Total utility** is a measure of the total satisfaction derived from consuming a given amount of goods and services.

Neal's total utility from each activity in this example is independent of the amount of the other activity he engages in. These total utilities are plotted in Figures 6.1 and 6.2. Clearly, more of each activity yields more utility, so the additional or **marginal utility** (MU) of each activity is positive. This positive marginal utility for any amount of the good consumed, no matter how much, reflects the assumption of *non-satiation*—more is always better. Note, however, that the decreasing slopes of the total utility curves show that *total utility is increasing at a diminishing rate*. While more is certainly better, each additional visit to Whistler or a jazz club augments Neal's utility by a smaller amount. At the margin, his additional utility declines: He has **diminishing marginal utility**. The marginal utilities associated with snowboarding and jazz are entered in columns 4 and 5 of Table 6.1. They are the differences in total utility values when consumption increases by one unit. For example, when Neal makes a sixth visit to Whistler his total utility increases from 260 utils to 292 utils. His marginal utility for the sixth unit is therefore 32 utils, as defined in column 4. In light of this example, it should be clear that we can define marginal utility as:

$$\text{Marginal Utility} = \frac{\text{additional utility}}{\text{additional consumption}} \text{ or, MU} = \frac{\Delta U}{\Delta C}, \tag{6.1}$$

where ΔC denotes the change in the quantity consumed of the good or service in question.

Marginal utility is the addition to total utility created when one more unit of a good or service is consumed.

Diminishing marginal utility implies that the addition to total utility from each extra unit of a good or service consumed is declining.

Figure 6.1: TU from snowboarding

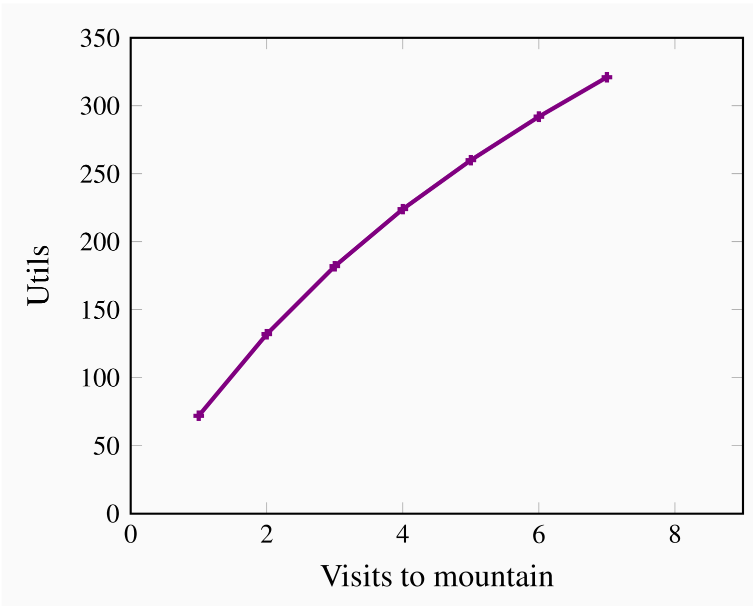


Figure 6.2: TU from jazz

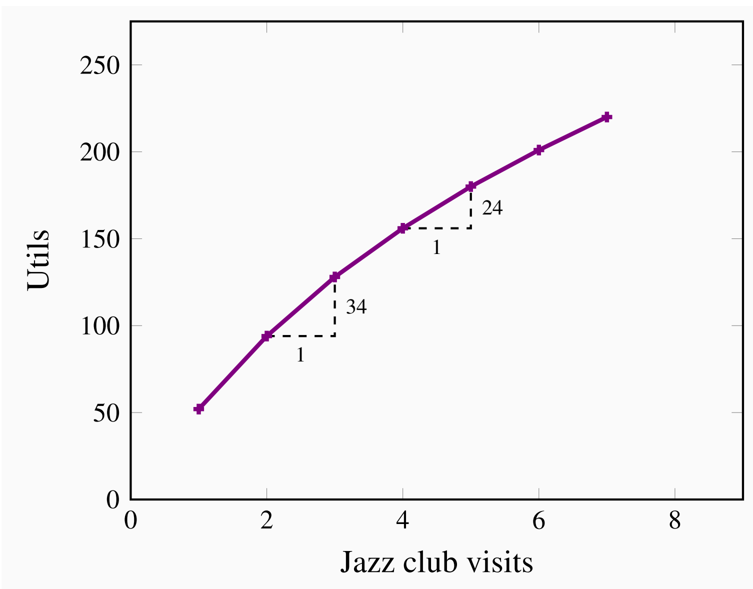


Figure 6.3: MU from snowboarding

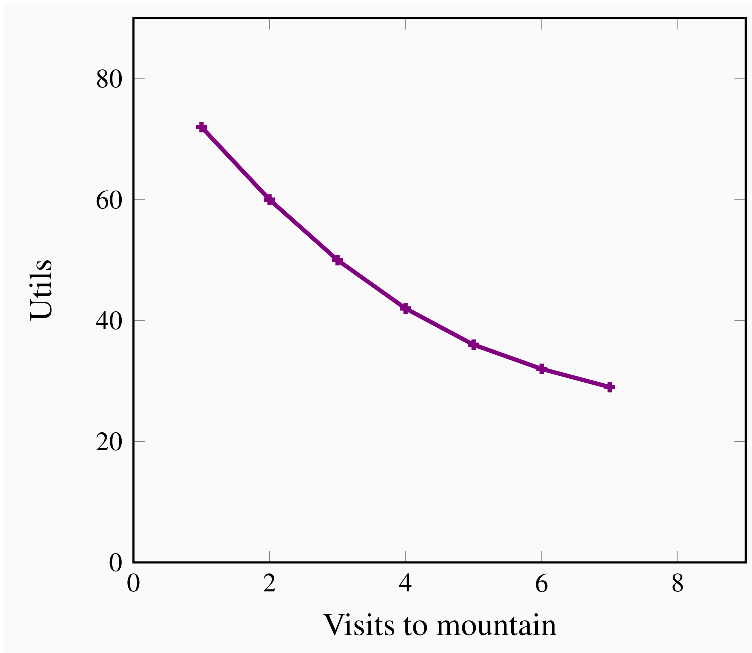
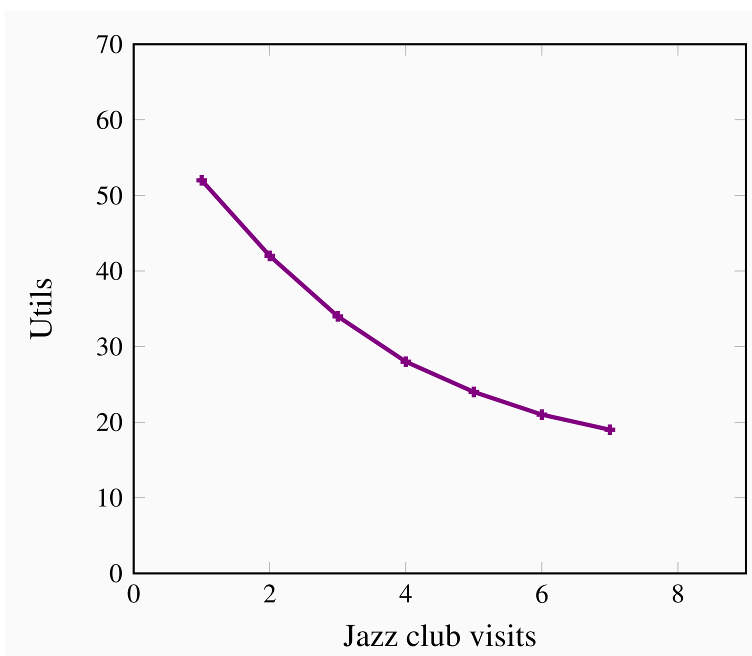


Figure 6.4: MU from jazz



The marginal utilities associated with consuming different amounts of the two goods are plotted in Figures 6.3 and 6.4, using the data from columns 4 and 5 in Table 6.1. These functions are declining, as indicated by their negative slope. It should also be clear that the *MU* curves can be derived from the TU curves. For example, in figure 6.2, when going from 2 units to 3 units of Jazz, TU increases by 34 units. But $34/1$ is the slope of the TU function in this range of consumption – the vertical distance divided by the horizontal distance. Similarly, if jazz consumption increases from 4 units to 5 units the corresponding change in TU is 24 units, again the vertical distance divided by the horizontal distance, and so the slope of the function. In short, the MU is the slope of the TU function.

Now that Neal has defined his utility schedules, he must consider the price of each activity. Ultimately, when deciding how to allocate his monthly entertainment budget, he must evaluate how much utility he gets from each dollar spent on snowboarding and jazz: What "bang for his buck" does he get? Let us see how he might go about allocating his budget. *When he has fully spent his budget in the manner that will yield him greatest utility, we say that he has attained equilibrium*, because he will have no incentive to change his expenditure patterns.

If he boards once, at a cost of \$30, he gets 72 utils of satisfaction, which is 2.4 utils per dollar spent ($= 72/30$). One visit to a jazz club would yield him 2.6 utils per dollar ($= 52/20$). Initially, therefore, his dollars give him more *utility per dollar* when spent on jazz. His MU per dollar spent on each activity is given in the final two columns of the table. These values are obtained by dividing the MU associated with each additional unit by the good's price.

We will assume that Neal has a budget of \$200. He realizes that his initial expenditure should be on a jazz club visit, because he gets more utility per dollar spent there. Having made one such expenditure, he sees that a second jazz outing would yield him 2.1 utils per dollar expended, while a first visit to Whistler would yield him 2.4 utils per dollar. Accordingly, his second activity is a snowboard outing.

Having made one jazz and one snowboarding visit, he then decides upon a second jazz club visit for the same reason as before—utility value for his money. He continues to allocate his budget in this way until his budget is exhausted. In our example, this occurs when he spends \$120 on four snowboarding outings and \$80 on four jazz club visits. At this **consumer equilibrium**, he gets the same utility value per dollar for the last unit of each activity consumed. This is a necessary condition for him to be maximizing his utility, that is, to be in equilibrium.

> **Consumer equilibrium** occurs when marginal utility per dollar spent on the last unit of each good is equal.

To be absolutely convinced of this, imagine that Neal had chosen instead to board twice and to visit the jazz clubs seven times; this combination would also exhaust his \$200 budget exactly. With such an allocation, he would get 2.0 utils per dollar spent on his marginal (second) snowboard outing, but just 0.95 utils per dollar spent on his marginal (seventh) jazz club visit.[1] If, instead, he were to reallocate his budget in favour of snowboarding, he would get 1.67 utils per dollar spent on

[1] Note that, with two snowboard outings and seven jazz club visits, total utility is 352 ($= 132 + 220$), while the optimal combination of four of each yields a total utility of 380 ($= 224 + 156$).

a third visit to the hills. By reducing the number of jazz visits by one, he would lose 0.95 utils per dollar reallocated. Consequently, the utility gain from a reallocation of his budget towards snowboarding would outweigh the utility loss from allocating fewer dollars to jazz. His initial allocation, therefore, was not an optimum, or equilibrium.

Only when the utility per dollar expended on each activity is equal at the margin will Neal be optimizing. When that condition holds, a reallocation would be of no benefit to him, because the gains from one more dollar on boarding would be exactly offset by the loss from one dollar less spent on jazz. Therefore, we can write the equilibrium condition as

$$\text{Equilibrium requires: } \frac{MU_s}{P_s} = \frac{MU_j}{P_j} \text{ or } \frac{MU_s}{MU_j} = \frac{P_s}{P_j}. \tag{6.2}$$

While this example has just two goods, in the more general case of many goods, this same *condition must hold for all pairs of goods* on which the consumer allocates his or her budget.

From utility to demand

Utility theory is a useful way of analyzing how a consumer makes choices. But in the real world we do not observe a consumer's utility, either total or marginal. Instead, his or her behaviour in the marketplace is observed through the demand curve. How are utility and demand related?

Demand functions relate the quantity of a good consumed to the price of that good, other things being equal. So let us trace out the effects of a price change on demand, with the help of this utility framework. We will introduce a simplification here: Goods are divisible, or that they come in small packages relative to income. Think, for example, of kilometres driven per year, or liters of gasoline purchased. Conceptualizing things in this way enables us to imagine more easily experiments in which small amounts of a budget are allocated one dollar at a time. In contrast, in the snowboard/jazz example, we had to reallocate the budget in lumps of $30 or $20 at a time because we could not "fractionalize" these goods.

The effects of a price change on a consumer's demand can be seen through the condition that describes his or her equilibrium. If income is allocated to, say, three goods $\{a,b,c\}$, such that $MU_a/P_a = MU_b/P_b = MU_c/P_c$, and the price of, say, good b falls, the consumer must reallocate the budget so that once again the MUs per dollar spent are all equated. How does he do this? Clearly, if he purchases more or less of any one good, the MU changes. If the price of good b falls, then the consumer initially gets more utility from good *b for the last dollar he spends on it* (the denominator in the expression MU_b/P_b falls, and consequently the value of the ratio rises to a value greater than the values for goods a and c).

The consumer responds to this, in the first instance, by buying more of the cheaper good. He obtains more total utility as a consequence, and in the process will get *less* utility *at the margin* from that good. In essence, the numerator in the expression then falls, in order to realign it with the lower price. This equality also provides an underpinning for what is called the **law of demand**:

More of a good is demanded at a lower price. If the price of any good falls, then, in order for the equilibrium condition to be re-established, the MU of that good must be driven down also. Since MU declines when more is purchased, this establishes that demand curves must slope downwards.

> The **law of demand** states that, other things being equal, more of a good is demanded the lower is its price.

However, the effects of a price decline are normally more widespread than this, because the quantities of other goods consumed may also change. As explained in earlier chapters, the decline in the price of good b will lead the consumer to purchase more units of *complementary goods* and fewer units of goods that are *substitutes*. So the whole budget allocation process must be redetermined in response to any price change. But at the end of the day, a new equilibrium must be one where the marginal utility per dollar spent on each good is equal.

Applying the theory

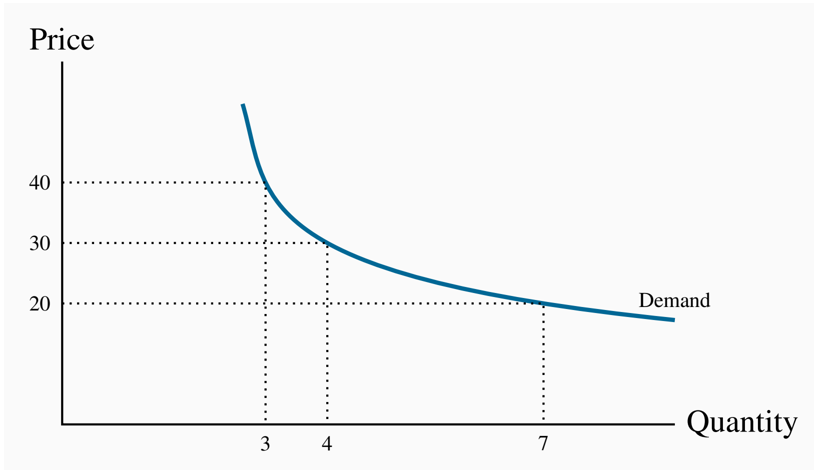
The demand curves developed in Chapter 3 can be related to the foregoing utility analysis. In our example, Neal purchased four lift tickets at Whistler when the price was \$30. We can think of this combination as one point on his demand curve, where the "other things kept constant" are the price of jazz, his income, his tastes, etc.

Suppose now that the price of a lift ticket increased to \$40. How could we find another point on his demand curve corresponding to this price, using the information in Table 6.1? The marginal utility per dollar associated with each visit to Whistler could be recomputed by dividing the values in column 4 by 40 rather than 30, yielding a new column 6. We would then determine a new allocation of his budget between the two goods that would maximize utility. After such a calculation we would find that he makes three visits to Whistler and four jazz-club visits. Thus, the combination $(P_s = \$40, Q_s = 3)$ is another point on his demand curve. Note that this allocation exactly exhausts his \$200 budget.

By setting the price equal to \$20, this exercise could be performed again, and the outcome will be a quantity demanded of lift tickets equal to seven (plus three jazz club visits). Thus, the combination $(P_s = \$20, Q_s = 7)$ is another point on his demand curve. Figure 6.5 plots a demand curve going through these three points.

By repeating this exercise for many different prices, the demand curve is established. We have now linked the demand curve to utility theory.

Figure 6.5: Utility to demand



When $P = \$30$, the consumer finds the quantity such that MU/P is equal for all purchases. The corresponding quantity purchased is 4 tickets. At prices of \$40 and \$20 the equilibrium condition implies quantities of 3 and 7 respectively.

Application Box 6.2: Individual and Collective Utility

The example developed in the text is not far removed from what economists do in practice. From a philosophical standpoint, economists are supposed to be interested in the well-being of the citizens who make up an economy or a country. To determine how 'well-off' citizens may be, social scientists frequently carry out surveys on how 'content' or 'happy' people are in their every-day lives. For example, the *Earth Institute at Columbia University* regularly produces a 'World Happiness Report'. The report is based upon responses to survey questions in numerous economies. One of the measures it uses to compare utility levels is the *Cantril ladder*. This is an 11-point scale running from 0 to 10, with the lowest value signifying the worst possible life, and 10 the highest possible quality of life. In reporting their findings, the researchers are essentially claiming that some economies have, on average, more contented or happier, people than others. Utility can be considered in exactly this way: A higher reported value on the Cantril ladder suggests higher utility.

A slightly different measure of well-being across economies is given by the *United Nations Human Development Index*. In this case, countries score high by having a high level of income, good health (as measured by life expectancy), and high levels of education, as measured by the number of years of education completed or envisaged.

In practice, social scientists are very comfortable using utility-based concepts to describe the economic circumstances of individuals in different economies.

6.3 Choice with ordinal utility

The budget constraint

In the preceding section, we assumed that utility is measurable in order to better understand how consumers allocate their budgets, and how this process is reflected in the market demands that are observed. The belief that utility might be measurable is not too extreme in the modern era. Neuroscientists are mapping more and more of the human brain and understanding how it responds to positive and negative stimuli. At the same time, numerous sociological surveys throughout the world ask individuals to rank their happiness on a scale of one to ten, or something similar, with a view to making comparisons between individual-level and group-level happiness – see Application Box 6.2. Nonetheless, not every scientist may be convinced that we should formulate behavioural rules on this basis. Accordingly we now examine the economics of consumer behaviour without this strong assumption. We assume instead that individuals are able to identify (a) different combinations of goods and services that yield equal satisfaction, and (b) combinations of goods and services that yield more satisfaction than other combinations. In contrast to measurable (or cardinal) utility, this concept is called ordinal utility, because it assumes only that consumers can order utility bundles rather than quantify the utility.

> **Ordinal utility** assumes that individuals can rank commodity bundles in accordance with the level of satisfaction associated with each bundle.

The budget constraint

Neal's monthly expenditure limit, or **budget constraint**, is \$200. In addition, he faces a price of \$30 for lift tickets and \$20 per visit to jazz clubs. Therefore, using S to denote the number of snowboard outings and J the number of jazz club visits, if he spends his entire budget it must be true that the sum of expenditures on each activity exhausts his budget or income (I):

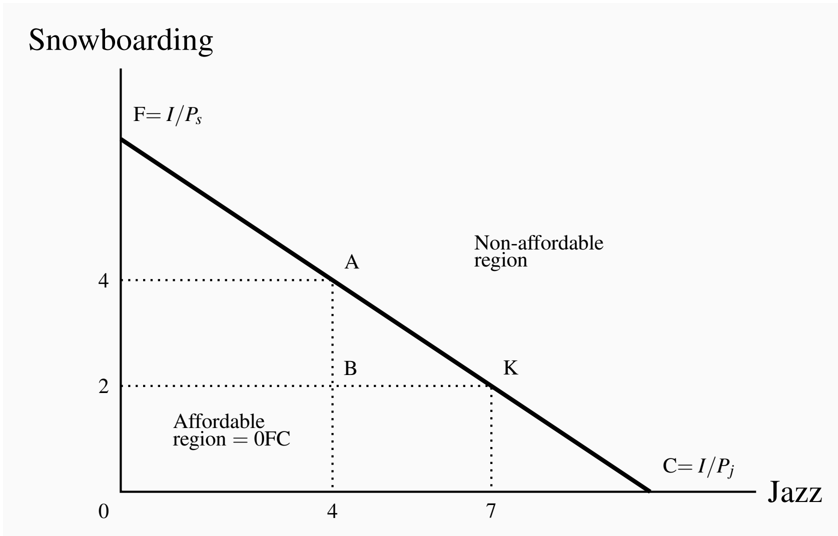
$$\begin{aligned}\text{Expenditure on snowboarding} + \text{expenditure on Jazz} &= \text{Income} \\ (\text{Price of } S \times \text{quantity of } S) + (\text{price of } J \times \text{quantity of } J) &= \text{Income} \\ P_S S + P_J J = I \ \text{ or } \ \$30S + \$20J &= \$200\end{aligned}$$

Since many different combinations of the two goods are affordable, it follows that the budget constraint defines all bundles of goods that the consumer can afford with a given budget.

> The **budget constraint** defines all bundles of goods that the consumer can afford with a given budget.

The budget constraint, then, is just what it claims to be—a limit on behaviour. Neal's budget constraint is illustrated in Figure 6.6, where the amount of each good consumed is given on the axes. If he spends all of his \$200 income on jazz, he can make exactly ten jazz club visits ($\$200/\$20 = 10$). The calculation also applies to visits to Whistler. The intercept value is always obtained by dividing income by the price of the good or activity in question.

Figure 6.6: The budget line



FC is the budget constraint and defines the affordable combinations of snowboarding and jazz. F represents all income spent on snowboarding. Thus F= I/P_s. Similarly C= I/P_j. Points above FC are not attainable. The slope = OF/OC = $(I/P_s)/(I/P_j) = P_j/P_s = 20/30 = 2/3$. The affordable set is 0FC.

In addition to these affordable extremes, Neal can also afford many other bundles, e.g., $(S = 2, J = 7)$, or $(S = 4, J = 4)$, or $(S = 6, J = 1)$. The set of feasible, or **affordable**, combinations is bounded by the budget line, and this is illustrated in Figure 6.6.

> The **affordable set** of goods and services for the consumer is bounded by the budget line from above; the **non-affordable set** lies strictly above the budget line.

The slope of the budget line is informative. As illustrated in Chapter 1, it indicates how many snowboard visits must be sacrificed for one additional jazz visit; it defines the consumer's *trade-offs*. To illustrate: Suppose Neal is initially at point A $(J = 4, S = 4)$, and moves to point K $(J = 7, S = 2)$. Clearly, both points are affordable. In making the move, he trades two snowboard outings in order to get three additional jazz club visits, a trade-off of 2/3. This trade-off is the slope of the budget line, which, in Figure 6.6, is AB/BK= $-2/3$, where the negative sign reflects the downward slope.

Could it be that this ratio reflects the two prices (\$20/\$30)? The answer is yes: The slope of the budget line is given by the vertical distance divided by the horizontal distance, OF/OC. The points F and C were obtained by dividing income by the respective price—remember that the jazz intercept is $\$200/\$20 = 10$. Formally, that is I/P_J. The intercept on the snowboard axis is likewise I/P_S. Accordingly, the slope of the budget constraint is:

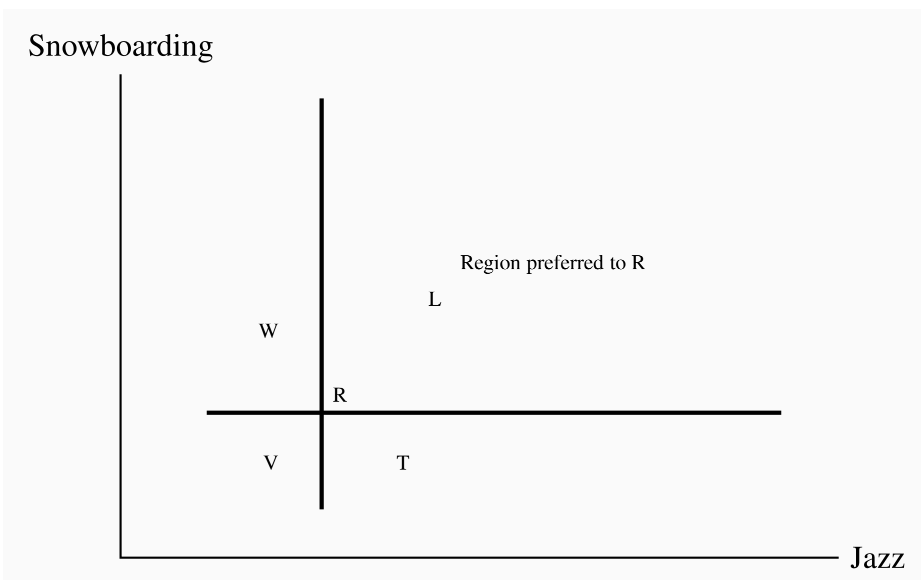
$$\text{Slope} = \text{OF/OC} = \frac{(I/P_s)}{(I/P_j)} = \frac{I}{P_s} \times \frac{P_j}{I} = \frac{P_j}{P_s}.$$

Since the budget line has a negative slope, it is technically correct to define it with a negative sign. But, as with elasticities, the sign is frequently omitted.

Tastes and indifference

We now consider how to represent a consumer's tastes in two dimensions, given that he can order, or rank, different consumption bundles, and that he can define a series of different bundles that all yield the same satisfaction. We limit ourselves initially to considering just "goods," and not "bads" such as pollution.

Figure 6.7: Ranking consumption bundles



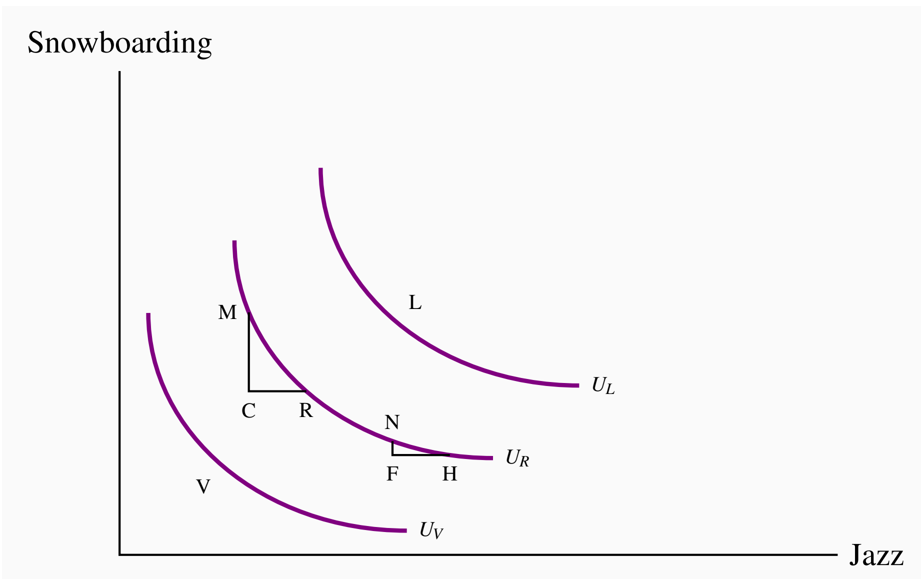
L is preferred to R since more of each good is consumed at L, while points such as V are less preferred than R. Points W and T contain more of one good and less of the other than R. Consequently, we cannot say if they are preferred to R without knowing how the consumer trades the goods off – that is, his preferences.

Figure 6.7 examines the implications of these assumptions about tastes. Each point shows a consumption bundle of snowboarding and jazz. Let us begin at bundle R. Since more of a good is preferred to less, any point such as L, which lies to the northeast of R, is preferred to R, since L offers more of both goods than R. Conversely, points to the southwest of R offer *less of each good* than R, and therefore R is preferred to a point such as V.

Without knowing the consumer's tastes, we cannot be sure at this stage how points in the northwest and southeast regions compare with R. At W or T, the consumer has more of one good and less of the other than at R. Someone who really likes snowboarding might prefer W to R, but a jazz buff might prefer T to R.

Let us now ask Neal to disclose his tastes, by asking him to define several combinations of snowboarding and jazz that *yield him exactly the same degree of satisfaction as the combination at R*. Suppose further, for reasons we shall understand shortly, that his answers define a series of points that lie on the beautifully smooth contour U_R in Figure 6.8. Since he is indifferent between all points on U_R by construction, this contour is an **indifference curve**.

Figure 6.8: Indifference curves



An indifference curve defines a series of consumption bundles, all of which yield the same satisfaction. The slope of an indifference curve is the marginal rate of substitution (*MRS*) and defines the number of units of the good on the vertical axis that the individual will trade for one unit of the good on the horizontal axis. The *MRS* declines as we move south-easterly, because the consumer values the good more highly when he has less of it.

An **indifference curve** defines combinations of goods and services that yield the same level of satisfaction to the consumer.

Pursuing this experiment, we could take other points in Figure 6.8, such as L and V, and ask the consumer to define bundles that would yield the same level of satisfaction, or indifference. These combinations would yield additional contours, such as U_L and U_V in Figure 6.8. This process yields a series of indifference curves that together form an **indifference map**.

> An **indifference map** is a set of indifference curves, where curves further from the origin denote a higher level of satisfaction.

Let us now explore the properties of this map, and thereby understand why the contours have their smooth convex shape. They have four properties. The first three follow from our preceding discussion, and the fourth requires investigation.

1. Indifference curves *further from the origin reflect higher levels of satisfaction*.
2. Indifference curves are *negatively sloped*. This reflects the fact that if a consumer gets more of one good she should have less of the other in order to remain indifferent between the two combinations.
3. Indifference curves *cannot intersect*. If two curves were to intersect at a given point, then we would have two different levels of satisfaction being associated with the same commodity bundle—an impossibility.
4. Indifference curves are convex when viewed from the origin, reflecting a *diminishing marginal rate of substitution*.

The convex shape reflects an important characteristic of preferences: When consumers have a lot of some good, they value a marginal unit of it less than when they have a small amount of that good. More formally, they have a *higher marginal valuation at low consumption levels*—that first cup of coffee in the morning provides greater satisfaction than the second or third cup.

Consider the various points on U_R, starting at M in Figure 6.8. At M Neal snowboards a lot; at N he boards much less. The convex shape of his indifference map shows that he values a marginal snowboard trip more at N than at M. To see this, consider what happens as he moves along his indifference curve, starting at M. We have chosen the coordinates on U_R so that, in moving from M to R, and again from N to H, the additional amount of jazz is the same: CR=FH. From M, if Neal moves to R, he consumes an additional amount of jazz, CR. By definition of the indifference curve, he is willing to give up MC snowboard outings. The ratio MC/CR defines his willingness to substitute one good for the other. This ratio, being a vertical distance divided by a horizontal distance, is the slope of the indifference curve and is called the **marginal rate of substitution**, *MRS*.

> The **marginal rate of substitution** is the slope of the indifference curve. It defines the amount of one good the consumer is willing to sacrifice in order to obtain a given increment of the other, while maintaining utility unchanged.

At N, the consumer is willing to sacrifice the amount NF of boarding to get the same additional amount of jazz. Note that, when he boards *less*, as at N, he is willing to give up less boarding than when he has a lot of it, as at M, in order to get the same additional amount of jazz. His willingness to substitute *diminishes* as he moves from M to N: The quantity NF is less than the quantity MC. In order to reflect this taste characteristic, the indifference curve has a **diminishing marginal rate of substitution**: A flatter slope as we move down along its surface.

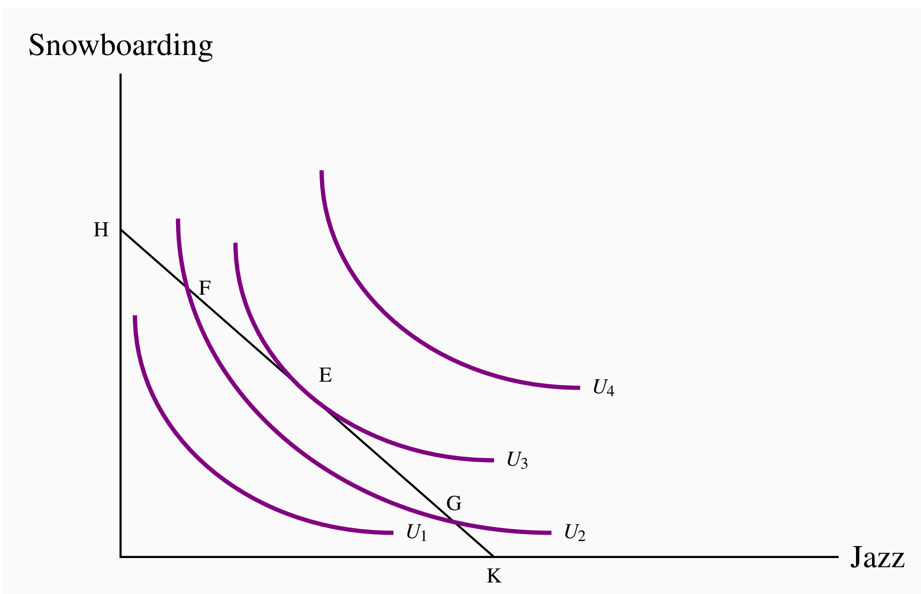
> A **diminishing marginal rate of substitution** reflects a higher marginal value being associated with smaller quantities of any good consumed.

Optimization

We are now in a position to examine how the consumer optimizes—how he gets to the highest level of satisfaction possible. The constraint on his behaviour is the affordable set defined in Figure 6.6, the budget line.

Figure 6.9 displays several of Neal's indifference curves in conjunction with his budget constraint. We propose that he maximizes his utility, or satisfaction, at the point E, on the indifference curve denoted by U_3. While points such as F and G are also on the boundary of the affordable set, they do not yield as much satisfaction as E, because E lies on a higher indifference curve. *The highest possible level of satisfaction is attained, therefore, when the budget line touches an indifference curve at just a single point—that is, where the constraint is tangent to the indifference curve*. E is such a point.

Figure 6.9: The consumer optimum



The budget constraint constrains the individual to points on or below HK. The highest level of satisfaction attainable is U_3, where the budget constraint just touches, or is just tangent to, it. At this optimum the slope of the budget constraint ($-P_j/P_s$) equals the MRS.

This tangency between the budget constraint and an indifference curve requires that the slopes of each be the same at the point of tangency. We have already established that the slope of the budget constraint is the negative of the price ratio ($= -P_x/P_y$). The slope of the indifference curve is the marginal rate of substitution *MRS*. It follows, therefore, that the **consumer optimizes** where the marginal rate of substitution equals the slope of the price line.

Optimization requires:

$$\text{Slope of Indifference curve} = \text{marginal rate of substitution} = -\frac{P_j}{P_s}. \qquad (6.3)$$

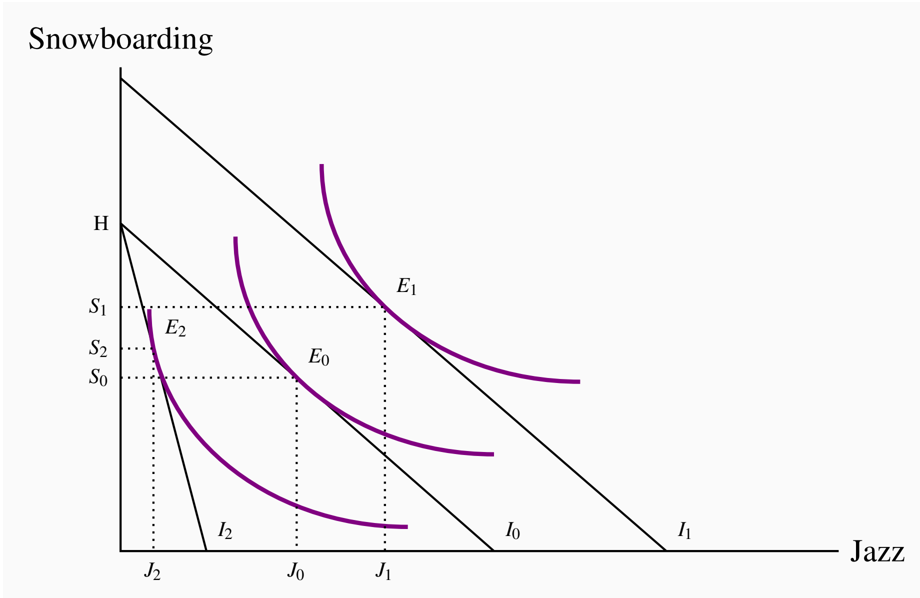
> A **consumer optimum** occurs where the chosen consumption bundle is a point such that the price ratio equals the marginal rate of substitution.

Notice the resemblance between this condition and the one derived in the first section as Equation 6.2. There we argued that equilibrium requires the ratio of the marginal utilities be same as the ratio of prices. Here we show that the *MRS* must equal the ratio of prices. In fact, with a little mathematics it can be shown that the *MRS* is indeed the same as the (negative of the) ratio of the marginal utilities: $MRS = -MU_j/MU_s$. Therefore the two conditions are in essence the same! However, it was not necessary to assume that an individual can actually measure his utility in obtaining the result that the *MRS* should equal the price ratio in equilibrium. The concept of ordinal utility is sufficient.

Adjusting to income changes

Suppose now that Neal's income changes from \$200 to \$300. How will this affect his consumption decisions? In Figure 6.10, this change is reflected in a *parallel* outward shift of the budget constraint. Since no price change occurs, the slope remains constant. By recomputing the ratio of income to price for each activity, we find that the new snowboard and jazz intercepts are 10 ($= \$300/\$30$) and 15 ($= \$300/\$20$), respectively. Clearly, the consumer can attain a higher level of satisfaction—at a new tangency to a higher indifference curve—as a result of the size of the affordable set being expanded. In Figure 6.10, the new equilibrium is at E_1.

Figure 6.10: Income and price adjustments



An income increase shifts the budget constraint from I_0 to I_1. This enables the consumer to attain a higher indifference curve. A price rise in jazz tickets rotates the budget line I_0 inwards around the snowboard intercept to I_2. The price rise reflects a lower real value of income and results in a lower equilibrium level of satisfaction.

Adjusting to price changes

Next, consider the impact of a price change from the initial equilibrium E_0 in Figure 6.10. Suppose that jazz now costs more. This reduces the purchasing power of the given budget of \$200. The new jazz intercept is therefore reduced. The budget constraint becomes steeper and rotates around the snowboard intercept H, which is unchanged because its price is constant. The new equilibrium is at E_2, which reflects a lower level of satisfaction because the affordable set has been reduced by the price increase. As explained in Section 6.2, E_0 and E_2 define points on the demand curve for jazz (J_0 and J_2): They reflect the consumer response to a change in the price of jazz with all other things held constant. In contrast, the price increase for jazz *shifts* the demand curve for snowboarding: As far as the demand curve for snowboarding is concerned, a change in the price of jazz is one of those things other than own-price that determine its position.

Philanthropy

Individuals in the foregoing analysis aim to maximize their utility, given that they have a fixed budget. Note that this behavioural assumption does not rule out the possibility that these same individuals may be philanthropic – that is, they get utility from the act of giving to their favourite charity or the United Way or Centre-aide. To see this suppose that donations give utility to the individual in question – she gets a 'warm glow' feeling as a result of giving, which is to say she gets utility from the activity. There is no reason why we cannot put charitable donations on one axis and some other good or combination of goods on the remaining axis. At equilibrium, the marginal utility per dollar of contributions to charity should equal the marginal utility per dollar of expenditure on other goods; or, stated in terms of ordinal utility, the marginal rate of substitution between philanthropy and any other good should equal the ratio of their prices. Evidently the price of a dollar of charitable donations is one dollar.

6.4 Applications of indifference analysis

Price impacts: Complements and substitutes

The nature of complements and substitutes, defined in Chapter 4, can be further understood with the help of Figure 6.10. The new equilibrium E_2 has been drawn so that the increase in the price of jazz results in more snowboarding—the quantity of S increases to S_2 from S_0. These goods are substitutes in this picture, because snowboarding *increases* in response to an *increase* in the price of jazz. If the new equilibrium E_2 were at a point yielding a lower level of S than S_0, we would conclude that they were complements.

Cross-price elasticities

Continuing with the same price increase in jazz, we could compute the *percentage* change in the quantity of snowboarding demanded as a result of the *percentage* change in the jazz price. In this example, the result would be a positive elasticity value, because the quantity change in snowboarding and the price change in jazz are both in the same direction, each being positive.

Income impacts: Normal and inferior goods

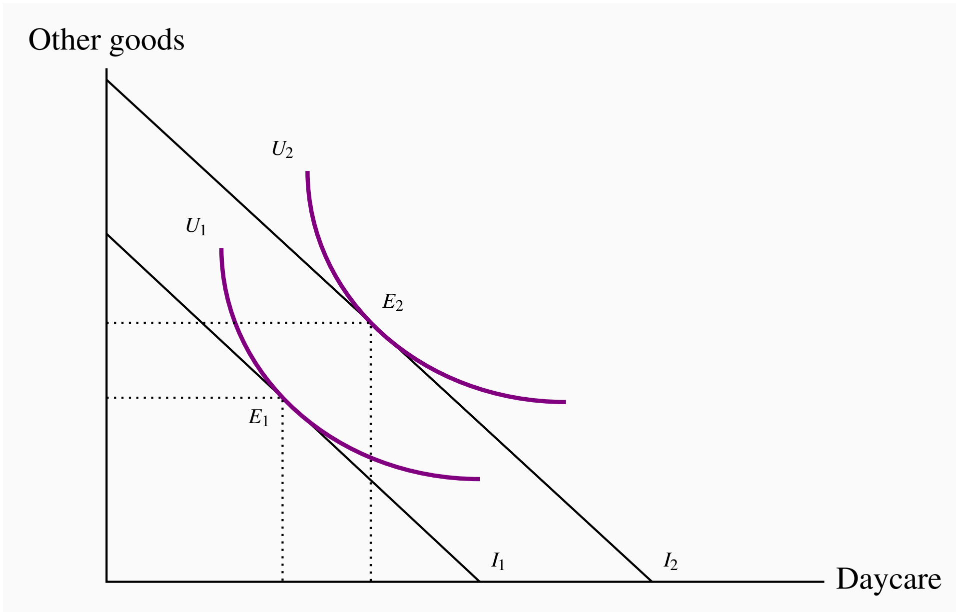
We know from Chapter 4 that the quantity demanded of a *normal good* increases in response to an income increase, whereas the quantity demanded of an *inferior good* declines. Clearly, both jazz and boarding are normal goods, as illustrated in Figure 6.10, because more of each one is demanded in response to the income increase from I_0 to I_1. It would challenge the imagination to think that either of these goods might be inferior. But if J were to denote junky (inferior) goods and S super goods, we could envisage an equilibrium E_1 to the northwest of E_0 in response to an income increase, along the constraint I_1; less J and more S would be consumed in response to the income increase.

Policy: Income transfers and price subsidies

Government policies that improve the purchasing power of low-income households come in two main forms: Pure income transfers and price subsidies. *Social Assistance* payments ("welfare") or *Employment Insurance* benefits, for example, provide an increase in income to the needy. Subsidies, on the other hand, enable individuals to purchase particular goods or services at a lower price—for example, rent or daycare subsidies.

In contrast to taxes, which *reduce* the purchasing power of the consumer, subsidies and income transfers *increase* purchasing power. The impact of an income transfer, compared with a pure price subsidy, can be analyzed using Figures 6.11 and 6.12.

Figure 6.11: Income transfer



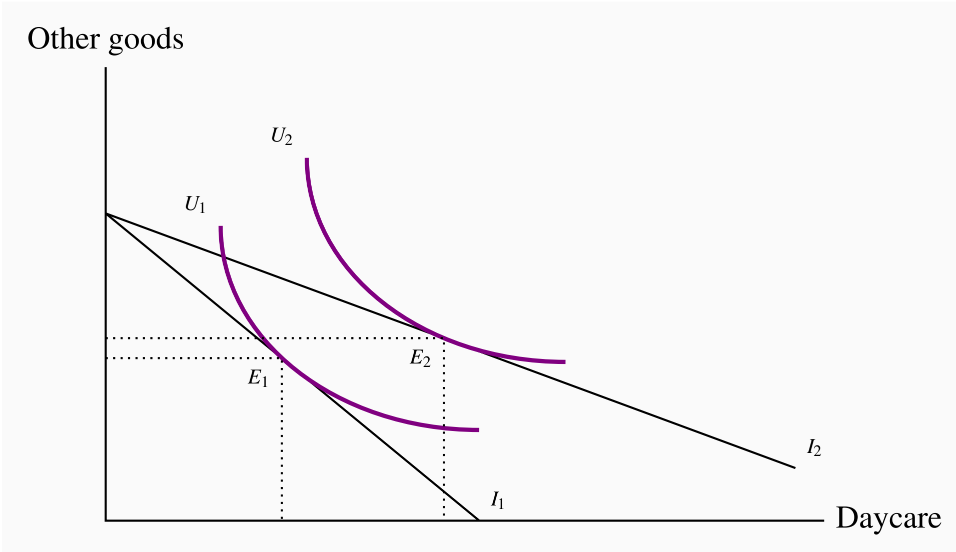
An increase in income due to a government transfer shifts the budget constraint from I_1 to I_2. This parallel shift increases the quantity consumed of the target good (daycare) *and* other goods, unless one is inferior.

In Figure 6.11, an *income transfer* increases income from I_1 to I_2. The new equilibrium at E_2 reflects an increase in utility, and an increase in the consumption of *both* daycare and other goods.

Suppose now that a government program administrator decides that, while helping this individual to purchase more daycare accords with the intent of the transfer, she does not intend that government money should be used to purchase other goods. She therefore decides that a daycare *subsidy* program might better meet this objective than a pure income transfer.

A daycare subsidy reduces the price of daycare and therefore *rotates the budget constraint outwards around the intercept on the vertical axis*. At the equilibrium in Figure 6.12, purchases of other goods change very little, and therefore most of the additional purchasing power is allocated to daycare.

Figure 6.12: Price subsidy



A subsidy to the targeted good, by reducing its price, rotates the budget constraint from I_1 to I_2. This induces the consumer to direct expenditure more towards daycare and less towards other goods than an income transfer that does not change the relative prices.

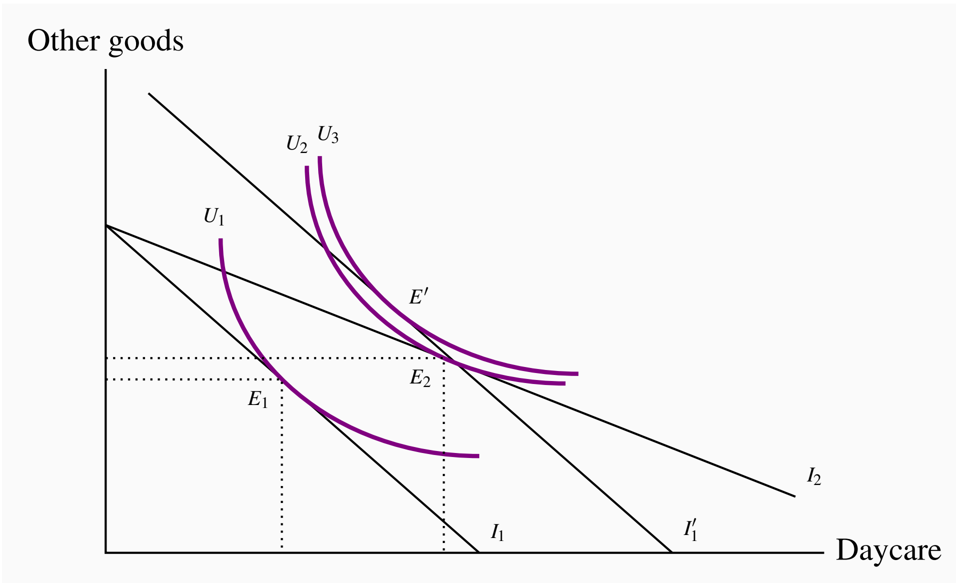
Let us take the example one stage further. From the initial equilibrium E_1 in Figure 6.12, suppose that, instead of a subsidy that took the individual to E_2, we gave an income transfer *that enabled the consumer to purchase the combination* E_2. Such a transfer is represented in Figure 6.13 by a parallel outward shift of the budget constraint from I_1 to I_1', going through the point E_2. We now have a subsidy policy and an alternative income transfer policy, each permitting the same consumption bundle (E_2). The interesting aspect of this pair of possibilities is that the income transfer will enable the consumer to attain a higher level of satisfaction—for example, at point E'—and will also induce her to consume more of the good on the vertical axis. The higher level of satisfaction comes about because the consumer has more latitude in allocating the additional real income.

Application Box 6.3: Daycare subsidies in Quebec

The Quebec provincial government subsidizes daycare heavily. In the public-sector network called the "Centres de la petite enfance", families can place their children in daycare for less than \$10 per day, while families that use the private sector are permitted a generous tax allowance for their daycare costs. This policy is designed to enable households to limit the share of their income expended on daycare. It is described in Figure 6.13.

The consequences of strong subsidization are not negligible: Excess demand, to such an extent that children are frequently placed on waiting lists for daycare places long before their parents intend to use the service. Annual subsidy costs amount to almost $2 billion per year. At the same time, it has been estimated that the policy has enabled many more parents to enter the workforce than otherwise would have.

Figure 6.13: Subsidy-transfer comparison



A price subsidy to the targeted good induces the individual to move from E_1 to E_2, facing a budget constraint I_2. An income transfer that permits him to consume E_2 is given by I_1'; but it also permits him to attain a higher level of satisfaction, denoted by E' on the indifference curve U_3.

The price of giving

Imagine now that the good on the horizontal axis is charitable donations, rather than daycare, and the government decides that for every dollar given the individual will see a reduction in their income tax of 50 cents. This is equivalent to cutting the 'price' of donations in half, because a donation of one dollar now costs the individual half of that amount. Graphically the budget constraint rotates outward with the vertical intercept unchanged. Since donations now cost less the individual has increased spending power as a result of the price reduction for donations. The price reduction is designed to increase the attractiveness of donations to the utility maximizing consumer.

Key Terms

Cardinal utility is a measurable concept of satisfaction.

Total utility is a measure of the total satisfaction derived from consuming a given amount of goods and services.

Marginal utility is the addition to total utility created when one more unit of a good or service is consumed.

Diminishing marginal utility implies that the addition to total utility from each extra unit of a good or service consumed is declining.

Consumer equilibrium occurs when marginal utility per dollar spent on the last unit of each good is equal.

Law of demand states that, other things being equal, more of a good is demanded the lower is its price.

Ordinal utility assumes that individuals can rank commodity bundles in accordance with the level of satisfaction associated with each bundle.

Budget constraint defines all bundles of goods that the consumer can afford with a given budget.

Affordable set of goods and services for the consumer is bounded by the budget line from above; the **non-affordable set** lies strictly above the budget line.

Indifference curve defines combinations of goods and services that yield the same level of satisfaction to the consumer.

Indifference map is a set of indifference curves, where curves further from the origin denote a higher level of satisfaction.

Marginal rate of substitution is the slope of the indifference curve. It defines the amount of one good the consumer is willing to sacrifice in order to obtain a given increment of the other, while maintaining utility unchanged.

Diminishing marginal rate of substitution reflects a higher marginal value being associated with smaller quantities of any good consumed.

Consumer optimum occurs where the chosen consumption bundle is a point such that the price ratio equals the marginal rate of substitution.

Exercises for Chapter 6

Exercise 6.1 In the example given in Table 6.1, suppose Neal experiences a small increase in income. Will he allocate it to snowboarding or jazz? [*Hint*: At the existing equilibrium, which activity will yield the higher MU for an additional dollar spent on it?]

Exercise 6.2 Suppose that utility depends on the square root of the amount of good X consumed: $U = \sqrt{X}$.

(a) In a spreadsheet enter the values $1 \ldots 16$ as the X column (col A), and in the adjoining column (B) compute the value of utility corresponding to each quantity of X. To do this use the 'SQRT' command. For example, the entry in cell B3 will be of the form '=SQRT(A3)'.

(b) In the third column enter the marginal utility (MU) associated with each value of X – the change in utility in going from one value of X to the next.

(c) Use the 'graph' tool to map the relationship between U and X.

(d) Use the graph tool to map the relationship between MU and X.

Exercise 6.3 Instead of the square-root utility function in Exercise 6.2, suppose that utility takes the form $U = x^2$.

(a) Follow the same procedure as in the previous question – graph the utility function.

(b) Why is this utility function not consistent with our beliefs on utility?

Exercise 6.4

(a) Plot the utility function $U = 2X$, following the same procedure as in the previous questions.

(b) Next plot the marginal utility values in a graph. What do we notice about the behaviour of the MU?

Exercise 6.5 Let us see if we can draw a utility function for beer. In this instance the individual may reach a point where he takes too much.

(a) If the utility function is of the form $U = 6X - X^2$, plot the utility values for X values in the range $1 \ldots 8$, using either a spreadsheet or manual calculations.

(b) At how many units of X (beer) is the individual's utility maximized?

(c) At how many beers does the utility become negative?

Exercise 6.6 Cappuccinos, C, cost \$3 each, and music downloads of your favourite artist, M, cost \$1 each from your iTunes store. Income is \$24.

(a) Draw the budget line, with cappuccinos on the vertical axis, and music on the horizontal axis, and compute the values of the intercepts.

(b) What is the slope of the budget constraint, and what is the opportunity cost of 1 cappuccino?

(c) Are the following combinations of goods in the affordable set: (4C and 9M), (6C and 2M), (3C and 15M)?

(d) Which combination(s) above lie inside the affordable set, and which lie on the boundary?

Exercise 6.7 George spends his income on gasoline and "other goods."

(a) First, draw a budget constraint, with gasoline on the horizontal axis.

(b) Suppose now that, in response to a gasoline shortage in the economy, the government imposes a *ration* on each individual that limits the purchase of gasoline to an amount less than the gasoline intercept of the budget constraint. Draw the new effective budget constraint.

Exercise 6.8 Suppose that you are told that the indifference curves defining the trade-off for two goods took the form of straight lines. Which of the four properties outlines in Section 6.3 would such indifference curves violate?

Exercise 6.9 Draw an indifference map with several indifference curves and several budget constraints corresponding to different possible levels of income. Note that these budget constraints should all be parallel because only income changes, not prices. Now find some optimizing (tangency) points. Join all of these points. You have just constructed what is called an income-consumption curve. Can you understand why it is called an income-consumption curve?

Exercise 6.10 Draw an indifference map again, in conjunction with a set of budget constraints. This time the budget constraints should each have a different price of good X and the same price for good Y.

(a) Draw in the resulting equilibria or tangencies and join up all of these points. You have just constructed a price-consumption curve for good X. Can you understand why the curve is so called?

(b) Now repeat part (a), but keep the price of X constant and permit the price of Y to vary. The resulting set of equilibrium points will form a price consumption curve for good Y.

Exercise 6.11 Suppose that movies are a normal good, but public transport is inferior. Draw an indifference map with a budget constraint and initial equilibrium. Now let income increase and draw a plausible new equilibrium, noting that one of the goods is inferior.

Chapter 7

Firms, investors and capital markets

In this chapter we will explore:

7.1 Business organization

Suppliers of goods and services to the marketplace come in a variety of forms; some are small, some are large. But, whatever their size, suppliers choose an organizational structure that is appropriate for their business: Aircraft, oil rigs, social media and information services are produced by large corporations; dental services and family health are provided by individual professionals or private partnerships.

The initial material of this chapter addresses organizational forms, their goals and their operation. We then examine why individuals choose to invest in firms, and illustrate that such investment provides individual investors with a means both to earning a return on their savings and to managing the risk associated with investing. Uncertainty regarding the future is a central consideration.

Understanding the way firms and capital markets function is crucial to understanding our economic history and how different forms of social and economic institutions interact. For example, seventeenth-century Amsterdam had a thriving bourgeoisie, well-developed financial markets, and investors with savings. This environment facilitated the channeling of investors' funds to firms specializing in trade and nautical conquest. This tiny state was then the source of some of the world's leading explorers and traders, and it had colonies stretching to Indonesia. The result was economic growth and prosperity.

In contrast, for much of the twentieth century, the Soviet Union dominated a huge territory covering much of Asia and Europe. But capital markets were non-existent, independent firms were stifled, and economic decline ultimately ensued. Much of the enormous difference in the respective patterns of economic development can be explained by the fact that one state fostered firms, capital markets, and legal institutions, while the other did not. In terms of our production possibility frontier: One set of institutional arrangements was conducive to expanding the possibilities; the other was not. Sustainable new businesses invariably require investors at an early point in the lifecycle of the business. Accordingly, financial and legal institutions that facilitate the flow of savings and financial investment into new enterprises perform a vital function in the economy.

Businesses, or firms, have several different forms. At the smallest scale, a business takes the form of a **sole proprietor** or sole trader who is the exclusive owner. A sole trader gets all of the revenues from the firm and incurs all of the costs. Hence he may make profits or be personally liable for the losses. In the latter case his business or even personal assets may be confiscated to cover debts. Personal bankruptcy may result.

> **Sole proprietor** is the single owner of a business.

If a business is to grow, **partners** may be required. Such partners can inject money in exchange for a share of future profits. Firms where trust is involved, such as legal or accounting firms, typically adopt this structure. A firm is given credibility when customers see that partners invest their own wealth in it.

> **Partnership**: a business owned jointly by two or more individuals, who share in the profits and are jointly responsible for losses.

In order to expand and grow, a firm will need cash, perhaps partners, and investors. Providers of family health and dental services rely primarily on human expertise, and therefore they need relatively little physical capital. Hence their cash start-up needs are limited. But firms that produce aircraft, or develop software and organizational systems, need vast amounts of money for capital investment; pharmaceuticals may need a billion dollars worth of research and development to bring a new drug to the marketplace; ride-sharing companies need billions in order to establish their business globally. Such businesses must form corporations – also known as companies. Not all corporations are public; some are privately held, but relatively few large corporations are not publicly traded.

Large organizations have several inherent advantages over small organizations when a high output level is required. Specialization in particular tasks leads to increased efficiency for production workers. At the same time, non-production workers can perform a multitude of different tasks. If a large corporation decided to contract out every task involved in bringing its product to market, the costs of such agreements would be prohibitively high. In addition, synergies can arise from teamwork. New ideas and better work flow are more likely to materialize when individuals work in close proximity than when working as isolated units, no matter how efficient they may be individually. A key aspect of such large organizations is that *they have a legal identity separate from the managers and owners*.

> **Corporation or company** is an organization with a legal identity separate from its owners that produces and trades.

The owners of a corporation are known as its **shareholders**, and their object is usually to make profits. There also exist non-profit corporations whose objective may be philanthropic. Since our focus is upon markets, we will generally assume that profits form the objective of a typical corporation. The profits that accrue to a corporation may be paid to the shareholders in the form of a **dividend**, or retained in the corporation for future use. When large profits (or losses) accrue the value of the corporation increases (or decreases), and this is reflected in the value of each share of the company. If the value of each share in the company increases (decreases) there is a **capital gain (loss)** to the owners of the shares – the shareholders. In any given year shareholders may receive a dividend and also obtain a capital gain (or loss). The sum of the dividend and capital gain represents the return to owning corporate stock in that year. When this sum is adjusted for inflation it is termed the **real return on corporate stock**

> **Shareholders** invest in corporations and therefore are the owners.
>
> **Dividends** are payments made from after-tax profits to company shareholders.
>
> **Capital gains (losses)** arise from the ownership of a corporation when an individual sells a share at a price higher (lower) than when the share was purchased.
>
> **Real return to corporate stock** is the inflation-adjusted sum of dividends and capital gain (or loss).

A key difference between a company and a partnership is that a company involves **limited liability**, whereas a partnership does not. Limited liability means that the liability of the company is limited to the value of the company's assets. Shareholders cannot be further liable for any wrongdoing on the part of the company. Accordingly, partnerships and sole traders normally insure themselves and their operations. For example, all specialist doctors carry malpractice insurance, and engineers insure themselves against error.

> **Limited liability** means that the liability of the company is limited to the value of the company's assets.

Corporations use capital, labour, and human expertise to produce a good, to supply a service, or to act as an intermediary. Corporations are required to produce an annual income statement that accurately describes the operation of the firm. An example is given in Table 7.1.

Table 7.1: The Regal Bank of Toronto, 2025

Total Revenue	$	32.0b
Net income post tax	$	4.80b
Shares outstanding		640m
Net income/share	$	7.50
Dividends/share	$	2.50
Share price	$	72.0
Market capitalization	$	46.08b

The data in Table 7.1 define the main financial characteristics of an imaginary bank: the Regal Bank of Toronto in the year 2025. "Net income post-tax" represents after-tax profits. There are 640 million shares outstanding, and thus each share could be attributed a profit of \$7.50 ($=$ \$4.80b $\div$ 640m). Of this amount, \$2.50 is distributed to shareholders in the form of dividends per share. The remainder is held by the Corporation in the form of **retained earnings** - to be used for future investment primarily. Each share traded at a price of \$72.00. Given that there were 640 million shares, the total market valuation of the corporation at that time stood at \$46.08 billion ($=$ 640m $\times$ \$72.00).

Such information is publicly available for a vast number of corporations at the 'finance' section of major search engines such as *Google* or *Yahoo*.

> **Retained earnings** are the profits retained by a company for reinvestment and not distributed as dividends.

In Canada, the corporate sector as a whole tends to hold on to more than half of after-tax profits in the form of retained earnings. However there exists considerable variety in the behaviour of corporations, and most firms establish a pattern of how profits are allocated between dividends and retained earnings. In the Table 7.1 example, one third of profits are distributed; yet some corporations have a no-dividend policy. In these latter cases the benefit to investing in a firm must come in the form of capital gain to the owners of the shares.

7.2 Profit

Ownership and corporate goals

As economists, we believe that profit maximization accurately describes a typical firm's objective. However, since large firms are not run by their owners but by their executives or agents, it is frequently hard for the shareholders to know exactly what happens within a company. Even the board of directors—the guiding managerial group—may not be fully aware of the decisions, strategies, and practices of their executives and managers. Occasionally things go wrong, sometimes as a result of managers deciding to follow their own interests rather than the interests of the company. In technical terms, the interests of the corporation and its shareholders might not be aligned with the interests of its managers. For example, managers might have a short horizon and take steps to increase their own income in the short term, knowing that they will move to another job before the long-term effects of their decisions impact the firm.

At the same time, the marketplace for the ownership of corporations exerts a certain discipline: If firms are not as productive or profitable as possible, they may become *subject to takeover* by other firms. Fear of such takeover can induce executives and boards to maximize profits.

The shareholder-manager relationship is sometimes called a **principal-agent relationship**, and it can give rise to a principal-agent problem. If it is costly or difficult to monitor the behaviour of an agent because the agent has additional information about his own performance, the principal may not know if the agent is working to achieve the firm's goals. This is the **principal-agent problem**.

> **Principal or owner**: delegates decisions to an agent, or manager.
>
> **Agent**: usually a manager who works in a corporation and is directed to follow the corporation's interests.
>
> **Principal-agent problem**: arises when the principal cannot easily monitor the actions of the agent, who therefore may not act in the best interests of the principal.

In an effort to deal with such a challenge, corporate executives frequently get bonuses or **stock options** that are related to the overall profitability of their firm. Stock options usually take the form of an executive being allowed to purchase the company's stock in the future – but at a price that is predetermined. If the company's profits do increase, then the price of the company's stock will reflect this and increase likewise. Hence the executive has an incentive to work with the objective of increasing profits because that will enable him to buy the company stock in the future at a lower price than it will be worth.

> **Stock option**: an option to buy the stock of the company at a future date for a fixed, predetermined price.

The threat of takeover and the structure of rewards, together, imply that the assumption of profit maximization is a reasonable one.

Application Box 7.1: The 'Sub-Prime' mortgage crisis: A principal-agent problem

With a decline in interest and mortgage rates in the early part of the twenty first century, many individuals believed they could afford to buy a house because the borrowing costs were lower than before. Employees and managers of lending companies believed likewise, and they structured loans in such a way as to provide an incentive to low-income individuals to borrow. These mortgage loans frequently enabled purchasers to buy a house with only a 5% down payment, in some cases even less, coupled with a repayment schedule that saw low repayments initially but higher repayments subsequently. The initial interest cost was so low in many of these mortgages that it was even lower than the 'prime' rate – the rate banks charge to their most prized customers.

The crisis that resulted became known as the 'sub-prime' mortgage crisis. In many cases loan officers got bonuses based on the total value of loans they oversaw, regardless of the quality or risk associated with the loan. The consequence was that they had the incentive to make loans to customers to whom they would not have lent, had these employees and managers been lending their own money, or had they been remunerated differently. The outcomes were disastrous for numerous lending institutions. When interest rates climbed, borrowers could not repay their loans. The construction industry produced a flood of houses that, combined with the sale of houses that buyers could no longer afford, sent housing prices through the floor. This in turn meant that recent house purchasers were left with negative value in their homes – the value of their property was less than what they paid for it. Many such 'owners' simply returned the keys to their bank, declared bankruptcy and walked away. Some lenders went bankrupt; some were bailed out by the government, others bought by surviving firms. This is a perfect example of the principal agent problem – the managers of the lending institutions and their loan officers did not have the incentive to act in the interest of the owners of those institutions.

The broader consequence of this lending practice was a financial collapse greater than any since the Depression of the nineteen thirties. Assets of the world's commercial and investment banks plummeted in value. Their assets included massive loans and investments both directly and indirectly to the real estate market, and when real estate values fell, so inevitably did the value of the assets based on this sector. Governments around the world had to buy up bad financial assets from financial institutions, or invest massive amounts of taxpayer money in these same institutions. Otherwise the world's financial system might have collapsed, with unknowable consequences.

Taxpayers and shareholders together bore the burden of this disastrous investment policy. Shareholders in many banks saw their shares drop in value to just a few percent of what they had been worth a year or two prior to the collapse.

Economic and accounting profit

Economists and accountants frequently differ in how they measure profits. An accountant stresses the financial flows of corporate activity; the economist is, in addition, concerned with opportunity cost. Imagine that Felicity has just inherited \$250,000 and decides to pursue her dream by opening a clothing boutique. She quits her job that pays her \$55,000 per annum, invests her inheritance in the purchase of a small retail space on the high street and launches her business. At the end of her first year she records $110,000 in clothing sales, which she purchased from the wholesaler for $50,000. She pays herself a salary of $35,000 and has no other accounting costs because she owns her physical capital – the store. Her accounting profit for the year is given by the margin returned between the buying and selling price of her clothing ($60,000) minus her incurred costs ($35,000) in salary. Her accounting profit is thus $25,000. Should she be content with this sum?

Felicity's economist friend, Prudence, informs Felicity that her enterprise is not returning a profit by economic standards. Prudence points out that Felicity could earn $55,000 as an alternative to working in her own store, hence there is an additional implicit cost of $20,000 to be considered, because Felicity only draws a salary of $35,000. Furthermore, Felicity has invested $250,000 in her business to avoid rent. But that sum, invested at the going interest rate of 4%, could earn her $10,000 per annum. That too is a foregone income stream so it is an implicit cost. Altogether, the additional implicit costs, not included in the accounting flows, amount to $30,000, and these implicit costs exceed the 'accounting profits'. Thus no economic profits are being made, because the economist includes implicit costs in her profit calculation. In economic terms Felicity would be better off by returning to her job and investing her inheritance. That strategy would generate an income of $65,000, as opposed to the income of $60,000 that she generates from the boutique – a salary of $35,000 plus an accounting profit of $25,000.

We can summarize this: **Accounting profit** is the difference between revenues and explicit costs. **Economic profit** is the difference between revenue and the sum of explicit and implicit costs. **Explicit costs** are the measured financial costs; **Implicit costs** represent the opportunity cost of the resources used in production.

> **Accounting profit**: is the difference between revenues and explicit costs.
>
> **Economic profit**: is the difference between revenue and the sum of explicit and implicit costs.
>
> **Explicit costs**: are the measured financial costs.
>
> **Implicit costs**: represent the opportunity cost of the resources used in production.

We will return to these concepts in the following chapters. Opportunity cost, or implicit costs, are critical in determining the long-run structure of certain sectors in the economy.

7.3 Risk and the investor

Firms cannot grow without investors. A successful firm's founder always arrives at a point where more investment is required if her enterprise is to expand. Frequently, she will not be able to secure a sufficiently large loan for such growth, and therefore must induce outsiders to buy shares in her firm. She may also realize that expansion carries risk, and she may want others to share in this risk. Risk plays a central role in the life of the firm and the investor. Most investors prefer to avoid risk, but are prepared to assume a limited amount of it if the anticipated rewards are sufficiently attractive.

An illustration of risk-avoidance is to be seen in the purchase of home insurance. Most home owners who run even a small risk of seeing their house burn down, being flooded, or damaged by a gas leak purchase insurance. By doing so they are avoiding risk. But how much are they willing to pay for such insurance? If the house is worth $500,000 and the probability of its being destroyed is one in one thousand in a given year then, using an averaging perspective, individuals should be willing to pay an insurance premium of $500 per annum. That insurance premium represents what actuaries call a 'fair' gamble: If the probability of disaster is one in one thousand, then the 'fair' premium should be one thousandth the value of the home that is being insured. If the insurance company insures millions of homes, then on average it will have to pay for the replacement of one house for every one thousand houses it insures each year. So by charging homeowners a price that exceeds $500 the insurer will cover not only the replacement cost of homes, but in addition cover her administrative costs and perhaps make a profit. Insurers operate on the basis of what we sometimes call the 'law of large numbers'.

In fact however, most individuals are willing to pay more than this 'fair' amount, and actually do pay more. If the insurance premium is $750 or $1,000 the home-owner is paying more than is actuarially 'fair', but a person who dislikes risk may be willing to pay such an amount in order to avoid the risk of being uninsured.

Our challenge now is to explain why individuals who purchase home insurance on terms that are less than actuarially 'fair' in order to avoid risk are simultaneously willing to invest their retirement savings into risky companies. Companies, like homes, are risky; while they may not collapse or implode in any given year, they can have good or bad returns in any given year. Corporate returns are inherently unpredictable and therefore risky. The key to understanding the willingness of risk-averse individuals to invest in risky firms is to be found in the pooling of risks.

7.4 Risk pooling and diversification

Silicon Valley: from angel investors to public corporation

Risky firms frequently succeed in attracting investment through the **capital market** in the modern economy. A typical start-up firm in the modern economy originates in the form of an idea. The developers of *Uber* got the idea of simplifying and streamlining the ride sharing business (sometimes called the taxi business). In its simplest form the inventors developed an App that would link users to drivers. The developers of *Airbnb* got the idea that spare accommodations in individual homes could be used to satisfy the needs of travellers. The founders developed an efficient means of putting potential renters/guests in communication with suppliers of rooms, houses and condominiums. *WeWork* was founded on the belief that workers and small corporations, particularly those in the technology sector, have immediate and changing space needs. The result is that *WeWork* provides flexible work space on a 'just in time' basis, frequently on a shared overhead basis. Each of these corporations acts as an intermediary, and sells intermediation services.

Typically the initial funding for new ideas comes from a couple of founding partners who develop a model or prototype of their software or their business. Following trials, the founders may approach potential investors for 'small' amounts that will fund expansion. Such investors are frequently 'angel' investors because they are a source of funding that makes the difference between expansion and death for the venture in question. If the venture shows promise the founders seek 'round A' funding, and this funding may come from **venture capitalists** who specialize in new ventures. Further evidence of possible success may result in 'round B' funding, and this frequently amounts to hundreds of millions of dollars.

Venture funds are managed by partners, or capitalists, with reputations for being better able than most to predict which start-ups will ultimately see profitability. These venture capitalists invest both their own funds and the funds of individuals who entrust their accumulated savings to the investing partnership.

But extremely high risk is associated with most start-ups. Funding frequently takes place in an environment where the venture has no revenue; merely a product in the course of development. Winners in the new economy are recognized and celebrated. Bill Gates' *Microsoft*, Jeff Bezos' *Amazon*, Steve Jobs *Apple*, and Larry Page and Sergey Brin's *Google* are corporate giants with valuations approach one trillion dollars. But Elizabeth Holmes' *Theranos*, once with an implicit valuation of several billion dollars has expired. *Theranos* hired several hundred employees with the aim of developing blood tests for scores of purposes using just a pin-prick of blood. But it was a failure, despite attracting hundreds of millions of dollars in investment. In the year 2019 Canada had dozens of cannabis-based firms listed on the Canadian Securities Exchange. None of these is earning a profit in 2020, some have negligible earnings, yet several have a value in excess of one billion dollars. It is highly improbable that all will survive.

> **Capital market**: a set of financial institutions that funnels financing from investors into bonds and stocks.

> **Venture capital**: investment in a business venture, where the ultimate outcome is highly unpredictable.

How can we reconcile the fact that, while these firms carry extraordinary uncertainty, investors are still willing to part with large sums of money to fund development? And the investors are not only billionaires with a good sense of the marketplace; private individuals who save for their retirement also invest in risky firms on the advice of their financial manager. Let us explore how and why.

Dealing with risk

Most (sensible) investors hold a **portfolio** of investments, which is a combination of different stocks and bonds. By investing in different stocks and bonds rather than concentrating in one single investment or type of investment, an individual diversifies her portfolio, which is to say she engages in **risk pooling**. Venture capital partnerships act in the same way. They recognize that their investments in start-ups will yield both complete failures and some roaring successes; the variation in their outcomes will exceed the variation in the outcomes of an investor who invests in 'mature' corporations.

> **Portfolio**: a combination of assets that is designed to secure an income from investing and to reduce risk.
>
> **Risk pooling**: Combining individual risks in such a way that the aggregate risk is reduced.

A rigorous theory underlies this "don't put all of your eggs in the one basket" philosophy. The essentials of diversification or pooling are illustrated in the example given in Table 7.2 below.[1] There are two risky stocks here: Natural Gas (NG) and technology (Tech). Each stock is priced at \$100, and over time it is observed that each yields a \$10 return in good times and \$0 in bad times. The investor has \$200 to invest, and each sector independently has a 50% probability ($p = 0.5$) of good or bad times. This means that each stock should yield a \$5 return *on average*: half of all outcomes will yield \$10 and half will yield zero. The challenge here is to develop an investment strategy that minimizes the risk for the investor.

At this point we need a specific working definition of **risk**. We define it in terms of how much variation a stock might experience in its returns from year to year. Each of NG and Tech have returns of either \$0 or \$10, with equal probability. But what if the Tech returns were either +\$20 or −\$10 with equal probability; or +\$40 or −\$30 with equal probability? In each of these alternative scenarios the average outcome remains the same: A positive average return of \$5. If the returns profile to NG remains unchanged we would say that Tech is a riskier stock (than NG) if its returns were defined by one of the alternatives here. Note that the average return is unchanged, and we are defining risk in terms of the greater spread in the possible returns around an unchanged average. The key to minimizing risk in the investor's portfolio lies in exploring how the variation in returns can be minimized by pooling risks.

> **Risk measurement**: A higher degree of risk is associated with increased variation in the possible returns around an unchanged mean return.

[1] A different form of risk management is defined by the idea of risk spreading. Imagine that an oil supertanker has to be insured and the owner approaches one insurer. In the event of the tanker being ship-wrecked the damage caused by the resulting oil spill would be catastrophic – both to the environment and the insurance company. In this instance the insurance company may not benefit from the law of large numbers – it may not be insuring thousands of tankers and therefore would find it difficult to balance the potential claims with the annual insurance premiums. As a consequence, an insurer will spread the potential cost among other insurers – this is called risk spreading. The world's major insurers, such as *Lloyd's of London*, have hundreds of syndicates who each take on a small proportion of a big risk. These syndicates may again choose to subdivide their share among others, until the big risk becomes widely spread. In this way it is possible to insure against almost any event or possibility, no matter how large.

Table 7.2: Investment strategies with risky assets

Strategy	Expected returns with probabilities		
\$200 in NG	220 ($p = 0.5$)		200 ($p = 0.5$)
\$200 in Tech	220 ($p = 0.5$)		200 ($p = 0.5$)
\$100 in each	220 ($p = 0.25$)	210 ($p = 0.5$)	200 ($p = 0.25$)

The outcomes from three different investment strategies are illustrated in Table 7.2. By investing all of her \$200 in either NG or Tech, she will obtain \$220 half of the time and \$200 half of the time, as indicated in the first two outcome rows. But by diversifying through buying one of each stock, as illustrated in the final row, she reduces the variability of her portfolio. To see why note that, since the performance of each stock is independent, there is now only a one chance in four that both stocks do well, and therefore there is a 25 percent probability of earning \$220. By the same reasoning, there is a 25 percent probability of earning \$200. But there is a 50 percent chance that one stock will perform well and the other poorly. When that happens, she gets a return of \$210. In contrast to the outcomes defined in rows 1 and 2, the **diversification** strategy in row 3 *yields fewer extreme potential outcomes and more potential outcomes that lie closer to the mean outcome*.

> **Diversification** reduces the total risk of a portfolio by pooling risks across several different assets whose individual returns behave independently.

Further diversification could reduce the variation in possible returns even further. To see this, imagine that, rather than having a choice between investing in one or two stocks, we could invest in four different stocks with the same returns profile as the two given in the table above. In such a case, the likelihood of getting extreme returns would be even lower than when investing in two stocks. This is because, if the returns to each stock are independent of the returns on the remaining stocks, it becomes increasingly improbable that all, or almost all, of the stocks will experience favorable (or unfavorable) returns in the same year. Now, imagine that we had 8 stocks, or 16, or 32, or 64, etc. The "magic" of diversification is that the same average return can be attained, yet variability can be reduced. If it can be reduced sufficiently by adding ever more stocks to the portfolio, then even a highly risk-averse individual can build a portfolio that is compatible with buying into risky firms.

We can conclude from this simple example that there need be no surprise over the fact that risk-averse individuals are willing, at the same time, to pay a high home-insurance premium to avoid risk, and simultaneously invest in risky ventures.

Application Box 7.2: The value of a financial advisor

The modern economy has thousands of highly-trained financial advisors. The successful ones earn huge salaries. But there is a puzzle: Why do such advisors exist? Can they predict the behaviour of the market any better than an uninformed advisor? Two insights help us answer the question.

First, Burton Malkiel wrote a best seller called *A Random Walk down Wall Street*. He provided ample evidence that a portfolio chosen on the basis of a monkey throwing darts at a list of stocks would do just as well as the average portfolio constructed by your friendly financial advisor.

Second, there are costs of transacting: An investor who builds a portfolio must devote time to the undertaking, and incur the associated financial trading cost. In recognizing this, investors may choose to invest in what they call mutual funds – a diversified collection of stocks – or may choose to employ a financial advisor who will essentially perform the same task of building a diversified portfolio. But, on average, financial advisors cannot beat the market, even though many individual investors would like to believe otherwise.

At this point we may reasonably ask why individuals choose to invest any of their funds in a "safe" asset – perhaps cash or Canadian Government bonds. After all, if their return to bonds is lower on average than the return to stocks, and they can diversify away much of the risk associated with stocks, why not get the higher average returns associated with stocks and put little or nothing in the safer asset? The reason is that it is impossible to fully diversify. When a recession hits, for example, the *whole stock market* may take a dive, because profits fall across the whole economy. On account of this possibility, we cannot ever arrive at a portfolio where the returns to the different stocks are completely independent. As a consequence, the rational investor will decide to put some funds in bonds in order to reduce this systematic risk component that is associated with the whole market. This is not a completely risk-free strategy because such assets can depreciate in value with inflation.

To see how the whole market can change dramatically students can go to any publicly accessible financial data site – such as *Yahoo Finance* and attempt to plot the TSX for the period 2005 – present, or the NASDAQ index from the mid-nineties to the present. The year 2020 is a particularly appropriate year to examine. In that year the coronavirus pandemic struck with disastrous impacts on stock markets worldwide. Initially almost all stocks declined in value. But within a matter of days investors realized that some firms would perform better in this particular downturn: those specializing in home delivery and those specializing in home exercise equipment for example. The stock valuations of firms such as Shopify, Amazon and Peloton shot up, while the valuations of traditional auto makers languished.

Efficiency and Allocation

We have now come full circle. We started this chapter by describing the key role in economic development and growth played by firms and capital markets. Capital markets channel the funds of individual investors to risk-taking firms. Such firms—whether they are Dutch spice importers in the seventeenth century, the Hudson's Bay Company in nineteenth-century Canada, communications corporations such as *Airbnb* or *Expedia*, or some high tech start-ups in Silicon Valley—are engines of growth and play a pivotal role in an economy's development. Capital markets are what make it possible for these firms to attract the savings of risk-averse individuals. By enabling individuals to diversify their portfolios, capital markets form the link between individuals and firms.

But capital markets fulfill another function, or at least they frequently do. They are a means of funnelling financial capital into ventures that appear to have a future return. It is not possible for each individual saver to perform the research necessary on a series of existing or new corporations, or 'ventures', in order to be able to invest in a knowledgeable manner. That is one reason we have financial intermediaries. When individuals deposit their savings with a bank, or with a financial manager, these individuals are anticipating that their savings will be protected and that a return will be forthcoming. A bank may promise a return of a fixed percent if an individual deposits her money in a guaranteed investment certificate. Alternatively, if the individual saver wishes to take on some risk she can place her savings with her financial manager, an equity fund, or even a venture capitalist. These intermediaries are better at assessing risks and returns than most private individuals. This in turn means that the return to the individual from entrusting their savings to one of them should on average exceed the returns that the individual would earn herself by following some investment strategy.

If the professional investor indeed invests in more profitable ventures than an amateur investor, then that intermediary is performing an efficiency function for the whole economy: He does better at directing the economy's savings to where it is more productive on a macro level. This in turn means that the economy should have a higher growth rate than if savings are allocated towards ventures that are less likely to grow and satisfy a need or a demand in the economy.

Consider the example of *Airbnb* that we cited earlier. The original intent of this corporation was to provide the owners of unused (home) space the opportunity to earn a return on that space. *Airbnb* was thus a transformative mechanism, in that it enabled unused resources to be more fully utilized - by linking potential buyers who were willing to pay for the product, with potential sellers who were willing to supply at a price buyers were willing to pay. Unused resources became utilized, created a surplus and contributed to growth in the macro economy.

While financial intermediaries perform a valuable service to both individual savers, and the economy at large, we should not expect that intermediaries always make optimal decisions. However, these analysts have research resources available, and thus they have a comparative advantage over individuals for whom investing is a part-time activity. By being more efficient than individuals, financial intermediaries perform their broader economic allocation function, even if that is an unintended by-product of their professional activity.

At times professional investors suffer from what has been called ‘irrational exuberance’. Crowd psychology creeps into the investment world from time to time, sometimes with devastating consequences. In the late 1990s tech stocks were all the rage and the stock market that specialized in trading such stocks saw the capital value of these stocks rise to stratospheric heights. The NASDAQ index stood at about 5,000 in March 2,000 but crashed to 1,300 by January of 2003. The run-up in NASDAQ valuations in the late nineties turned out to be a bubble.

Conclusion

We next turn to examine decision making *within* the firm. Firms must make the right decisions if they are to grow and provide investors with a satisfactory return. Firms that survive the growth process and ultimately bring a product to market are the survivors of the uncertainty surrounding product development.

Key Terms

Sole proprietor is the single owner of a business and is responsible for all profits and losses.

Partnership: a business owned jointly by two or more individuals, who share in the profits and are jointly responsible for losses.

Corporation or company is an organization with a legal identity separate from its owners that produces and trades.

Shareholders invest in corporations and therefore are the owners. They have limited liability personally if the firm incurs losses.

Dividends are payments made from after-tax profits to company shareholders.

Capital gains (losses) arise from the ownership of a corporation when an individual sells a share at a price higher (lower) than when the share was purchased.

Real return on corporate stock: the sum of dividend plus capital gain, adjusted for inflation.

Real return: the nominal return minus the rate of inflation.

Limited liability means that the liability of the company is limited to the value of the company's assets.

Retained earnings are the profits retained by a company for reinvestment and not distributed as dividends.

Principal or owner: delegates decisions to an agent, or manager.

Agent: usually a manager who works in a corporation and is directed to follow the corporation's interests.

Principal-agent problem: arises when the principal cannot easily monitor the actions of the agent, who therefore may not act in the best interests of the principal.

Stock option: an option to buy the stock of the company at a future date for a fixed, predetermined price.

Accounting profit: is the difference between revenues and explicit costs.

Economic profit: is the difference between revenue and the sum of explicit and implicit costs.

Explicit costs: are the measured financial costs.

Implicit costs: represent the opportunity cost of the resources used in production.

Capital market: a set of financial institutions that funnels financing from investors into bonds and stocks.

Portfolio: a combination of assets that is designed to secure an income from investing and to reduce risk.

Risk pooling: a means of reducing risk and increasing utility by aggregating or pooling multiple independent risks.

Risk: the risk associated with an investment can be measured by the dispersion in possible outcomes. A greater dispersion in outcomes implies more risk.

Diversification reduces the total risk of a portfolio by pooling risks across several different assets whose individual returns behave independently.

EXERCISES FOR CHAPTER 7

Exercise 7.1 Henry is contemplating opening a microbrewery and investing his savings of $100,000 in it. He will quit his current job as a quality controller at Megaweiser where he is paid an annual salary of $50,000. He plans on paying himself a salary of $40,000 at the microbrewery. He also anticipates that his beer sales minus all costs other than his salary will yield him a surplus of $55,000 per annum. The rate of return on savings is 7%.

(a) Calculate the accounting profits envisaged by Henry.

(b) Calculate the economic profits.

(c) Should Henry open the microbrewery?

(d) If all values except the return on savings remain the same, what rate of return would leave him indifferent between opening the brewery and not?

Exercise 7.2 You see an advertisement for life insurance for everyone 55 years of age and older. The advertisement says that no medical examination is required prior to purchasing insurance. If you are a very healthy 57-year old, do you think you will get a good deal from purchasing this insurance?

Exercise 7.3 In which of the following are risks being pooled, and in which would risks likely be spread by insurance companies?

(a) Insurance against Alberta's Bow River Valley flooding.

(b) Life insurance.

(c) Insurance for the voice of Avril Lavigne or Celine Dion.

(d) Insuring the voices of the lead vocalists in Metallica, Black Eyed Peas, Incubus, Evanescence, Green Day, and Jurassic Five.

Exercise 7.4 Your house has a one in five hundred probability chance of burning down in any given year. It is valued at $350,000.

(a) What insurance premium would be actuarially fair for this situation?

(b) If the owner is willing to pay a premium of $900, does she dislike risk or is she indifferent to risk?

Exercise 7.5 If individuals experience diminishing marginal utility from income it means that their utility function will resemble the total utility functions developed graphically in Section 6.2. Let us imagine specifically that if Y is income and U is utility, the individual gets utility from income according to the relation $U = \sqrt{Y}$.

(a) In a spreadsheet or using a calculator, calculate the amount of utility the individual gets for all income values running from \$1 to \$25.

(b) Graph the result with utility on the vertical axis and income on the horizontal axis, and verify from its shape that the marginal utility of income is declining.

(c) Using your calculations, how much utility will the individual get from \$4, \$9 and \$16?

(d) Suppose now that income results from a lottery and half of the time the individual gets \$4 and half of the time he gets \$16. How much utility will he get on average?

(e) Now suppose he gets \$10 each time with certainty. How much utility will he get from this?

(f) Since \$10 is exactly an average of \$4 and \$16, can you explain why \$10 with certainty gives him more utility than getting \$4 and \$16 each half of the time?

Exercise 7.6 In Question 7.5, suppose that the individual gets utility according to the relation $U = \frac{1}{2}Y$. Repeat the calculations for each part of the question and see if you can understand why the answers are different.

Chapter 8

Production and cost

In this chapter we will explore:

8.1 Efficient production

Firms that fail to operate efficiently seldom survive. They are dominated by their competitors because the latter produce more efficiently and can sell at a lower price. The drive for profitability is everywhere present in the modern economy. Companies that promise more profit, by being more efficient, are valued more highly on the stock exchange. For example: In July of 2015 *Google* announced that, going forward, it would be more attentive to cost management in its numerous research endeavours that aim to bring new products to the marketplace. This policy, put in place by the Company's new Chief Financial Officer, was welcomed by investors who, as a result, bought up the stock. The Company's stock increased in value by 16% in one day – equivalent to about $50 billion.

The remuneration of managers in virtually all corporations is linked to profitability. Efficient production, *a.k.a.* cost reduction, is critical to achieving this goal. In this chapter we will examine cost management and efficient production from the ground up – by exploring how a small entrepreneur brings his or her product to market in the most efficient way possible. As we shall see, efficient production and cost minimization amount to the same thing: Cost minimization is the financial reflection of efficient production.

Efficient production is critical in any budget-driven organization, not just in the private sector. Public institutions equally are, and should be, concerned with costs and efficiency.

Entrepreneurs employ factors of production (capital and labour) in order to transform raw materials and other inputs into goods or services. The relationship between output and the inputs used in the production process is called a **production function**. It specifies how much output can be produced with given combinations of inputs. A production function is not restricted to profit-driven organizations. Municipal road repairs are carried out with labour and capital. Students are educated with teachers, classrooms, computers, and books. Each of these is a production process.

> **Production function**: a technological relationship that specifies how much output can be produced with specific amounts of inputs.

Economists distinguish between two concepts of efficiency: One is **technological efficiency**; the other is **economic efficiency**. To illustrate the difference, consider the case of auto assembly: the assembler could produce its vehicles either by using a large number of assembly workers and a plant that has a relatively small amount of machinery, or it could use fewer workers accompanied by more machinery in the form of robots. Each of these processes could be deemed technologically efficient, provided that there is no waste. If the workers without robots are combined with their capital to produce as much as possible, then that production process is technologically efficient. Likewise, in the scenario with robots, if the workers and capital are producing as much as possible, then that process too is efficient in the technological sense.

> **Technological efficiency** means that the maximum output is produced with the given set of inputs.

Economic efficiency is concerned with more than just technological efficiency. Since the entrepreneur's goal is to make profit, she must consider which technologically efficient process best achieves that objective. More broadly, any budget-driven process should focus on being economically efficient, whether in the public or private sector. An economically efficient production structure is the one that produces output at least cost.

> **Economic efficiency** defines a production structure that produces output at least cost.

Auto-assembly plants the world over have moved to using robots during the last two decades. Why? The reason is not that robots were invented 20 years ago; they were invented long before that. The real reason is that, until recently, this technology was not economically efficient. Robots were too expensive; they were not capable of high-precision assembly. But once their cost declined and their accuracy increased they became economically efficient. The development of robots represented technological progress. When this progress reached a critical point, entrepreneurs embraced it.

To illustrate the point further, consider the case of garment assembly. There is no doubt that engineers could make robots capable of joining the pieces of fabric that form garments. This is not beyond our technological abilities. Why, then, do we not have such capital-intensive production processes for garment making, similar to the production process chosen by vehicle producers? The answer is that, while such a concept could be technologically efficient, it would not be economically efficient. It is more profitable to use large amounts of labour and relatively traditional machines to assemble garments, particularly when labour in Asia costs less and the garments can be shipped back to Canada inexpensively. Containerization and scale economies in shipping mean that a garment can be shipped to Canada from Asia for a few cents per unit.

Efficiency in production is not limited to the manufacturing sector. Farmers must choose the optimal combination of labour, capital and fertilizer to use. In the health and education sectors, efficient supply involves choices on how many high- and low-skill workers to employ, how much traditional physical capital to use, how much information technology to use, based upon the productivity and cost of each. Professors and physicians are costly inputs. When they work with new technology (capital) they become more efficient at performing their tasks: It is less costly to have a single professor teach in a 300-seat classroom that is equipped with the latest technology, than have several professors each teaching 60-seat classes with chalk and a blackboard.

8.2 The time frame

We distinguish initially between the **short run** and the **long run**. When discussing technological change, we use the term **very long run**. These concepts have little to do with clocks or calendars; rather, they are defined by the degree of flexibility an entrepreneur or manager has in her production process. A key decision variable is capital.

A customary assumption is that a producer can hire more labour immediately, if necessary, either by taking on new workers (since there are usually some who are unemployed and looking for work), or by getting the existing workers to work longer hours. In contrast, getting new capital in place is usually more time consuming: The entrepreneur may have to place an order for new machinery, which will involve a production and delivery time lag. Or she may have to move to a more spacious location in order to accommodate the added capital. Whether this calendar time is one week, one month, or one year is of no concern to us. We define the long run as a period of sufficient length to enable the entrepreneur to adjust her capital stock, whereas in the short run at least one factor of production is fixed. Note that it matters little whether it is labour or capital that is fixed in the short run. A software development company may be able to install new capital (computing power) instantaneously but have to train new developers. In such a case capital is variable and labour is fixed in the short run. The definition of the short run is that one of the factors is fixed, and in our examples we will assume that it is capital.

> **Short run**: a period during which at least one factor of production is fixed. If capital is fixed, then more output is produced by using additional labour.
>
> **Long run**: a period of time that is sufficient to enable all factors of production to be adjusted.

Very long run: a period sufficiently long for new technology to develop.

8.3 Production in the short run

Black Diamond Snowboards (BDS) is a start-up snowboard producing enterprise. Its founder has invented a new lamination process that gives extra strength to his boards. He has set up a production line in his garage that has four workstations: Laminating, attaching the steel edge, waxing, and packing.

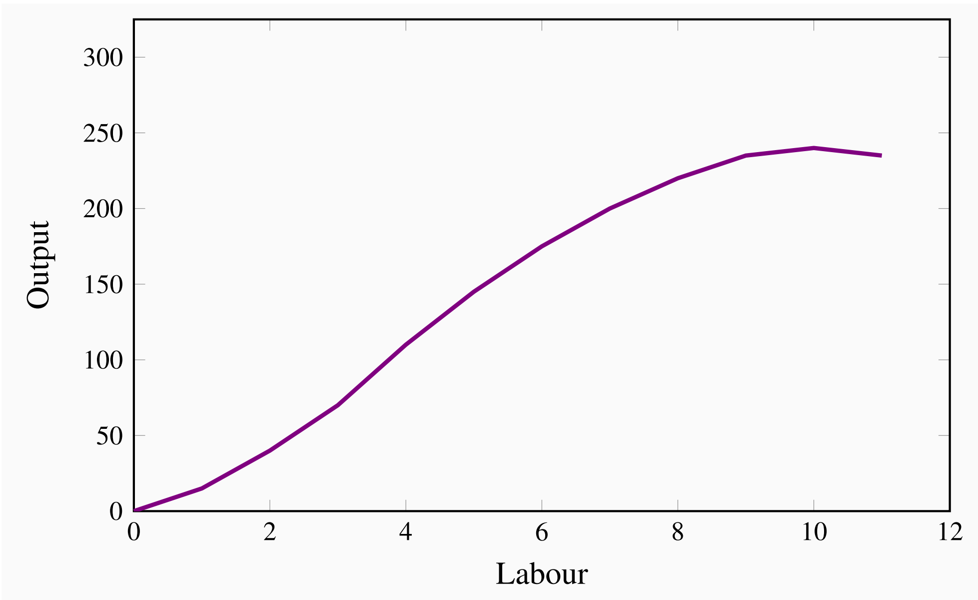
With this process in place, he must examine how productive his firm can be. After extensive testing, he has determined exactly how his productivity depends upon the number of workers. If he employs only one worker, then that worker must perform several tasks, and will encounter 'down time' between workstations. Extra workers would therefore not only increase the total output; they could, in addition, increase output *per worker*. He also realizes that once he has employed a critical number of workers, additional workers may not be so productive: Because they will have to share the fixed amount of machinery in his garage, they may have to wait for another worker to finish using a machine. At such a point, the productivity of his plant will begin to fall off, and he may want to consider capital expansion. But for the moment he is constrained to using this particular assembly plant. Testing leads him to formulate the relationship between workers and output that is described in Table 8.1.

Table 8.1: Snowboard production and productivity

1 **Workers**	**2** **Output** (TP)	**3** **Marginal product** (MP_L)	**4** **Average product** (AP_L)	**5** **Stages of production**
0	0			
1	15	15	15	
2	40	25	20	MP_L *increasing*
3	70	30	23.3	
4	110	40	27.5	
5	145	35	29	
6	175	30	29.2	
7	200	25	28.6	MP_L *positive and declining*
8	220	20	27.5	
9	235	15	26.1	
10	240	5	24.0	
11	235	-5	21.4	MP_L *negative*

By increasing the number of workers in the plant, BDS produces more boards. The relationship between these two variables in columns 1 and 2 in the table is plotted in Figure 8.1. This is called the **total product function** (TP), and it defines the output produced with different amounts of labour in a plant of fixed size.

Figure 8.1: Total product curve



Output increases with the amount of labour used. Initially the increase in output due to using more labour is high, subsequently it is lower. The initial phase characterizes increasing productivity, the later phase defines declining productivity.

> **Total product** is the relationship between total output produced and the number of workers employed, for a given amount of capital.

This relationship is positive, indicating that more workers produce more boards. But the curve has an interesting pattern. In the initial expansion of employment it becomes progressively steeper – its curvature is slightly convex; following this phase the function's increase becomes progressively less steep – its curvature is concave. These different stages in the TP curve tell us a great deal about productivity in BDS. To see this, consider the additional number of boards produced by each worker. The first worker produces 15. When a second worker is hired, the total product rises to 40, so the additional product attributable to the second worker is 25. A third worker increases output by 30 units, and so on. We refer to this additional output as the marginal product (MP) of an additional worker, because it defines the incremental, or marginal, contribution of the worker. These values are entered in column 3.

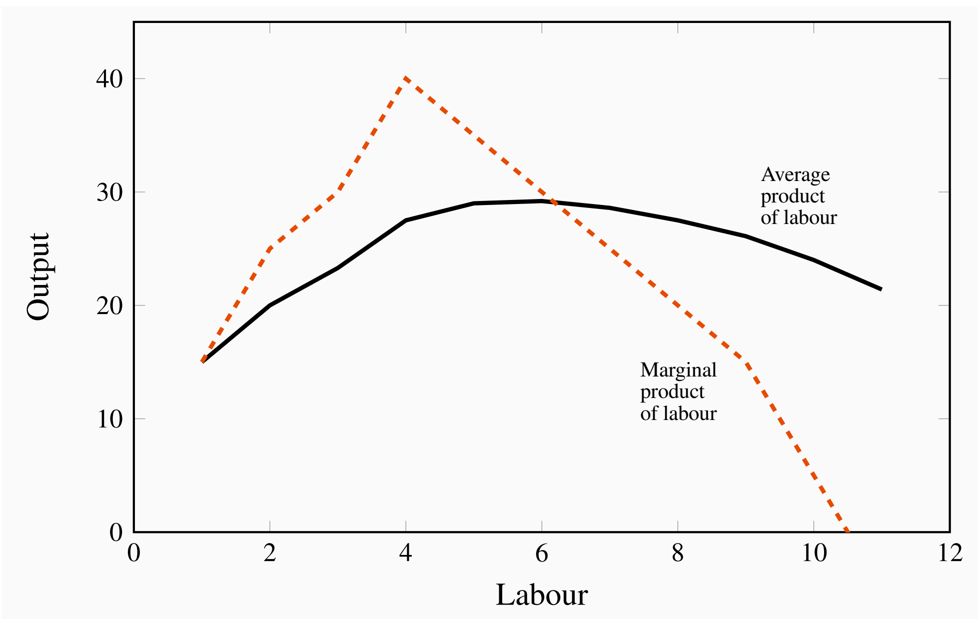
More generally the MP of labour is defined as the change in output divided by the change in the number of units of labour employed. Using, as before, the Greek capital delta (Δ) to denote a change, we can define

$$MP_L = \frac{\text{Change in output produced}}{\text{Change in labour employed}} = \frac{\Delta Q}{\Delta L}$$

In this example the change in labour is one unit at each stage and hence the **marginal product of labour** is simply the corresponding change in output. It is also the case that the MP_L is the slope of the TP curve – the change in the value on the vertical axis due to a change in the value of the variable on the horizontal axis.

> **Marginal product of labour** is the addition to output produced by each additional worker. It is also the slope of the total product curve.

Figure 8.2: Average and marginal product curves



The productivity curves initially rise and then decline, reflecting increasing and decreasing productivity. The MP_L curves must intersect the AP_L curve at the maximum of the AP_L: The average must increase if the marginal exceeds the average and must decline if the marginal is less than the average.

During the initial stage of production expansion, the marginal product of each worker is increasing. It increases from 15 to 40 as BDS moves from having one employee to four employees. This increasing *MP* is made possible by the fact that each worker is able to spend more time at his workstation, and less time moving between tasks. But, at a certain point in the employment expansion, the MP reaches a maximum and then begins to tail off. At this stage – in the concave region of the TP curve – additional workers continue to produce additional output, but at a diminishing rate. For example, while the fourth worker adds 40 units to output, the fifth worker adds 35, the sixth worker 30, and so on. This declining MP is due to the constraint of a fixed number of machines: All workers must share the same capital. The MP function is plotted in Figure 8.2.

The phenomenon we have just described has the status of a law in economics: The **law of diminishing returns** states that, in the face of a fixed amount of capital, the contribution of additional units of a variable factor must eventually decline.

> **Law of diminishing returns**: when increments of a variable factor (labour) are added to a fixed amount of another factor (capital), the marginal product of the variable factor must eventually decline.

The relationship between Figures 8.1 and 8.2 should be noted. First, the MP_L reaches a maximum at an output of 4 units – where the slope of the TP curve is greatest. The MP_L curve remains positive beyond this output, but declines: The TP curve reaches a maximum when the tenth unit of labour is employed. An eleventh unit actually reduces total output; therefore, the MP of this eleventh worker is negative! In Figure 8.2, the MP curve becomes negative at this point. The garage is now so crowded with workers that they are beginning to obstruct the operation of the production process. Thus the producer would never employ an eleventh unit of labour.

Next, consider the information in the fourth column of the table. It defines the average product of labour (AP_L)—the amount of output produced, on average, by workers at different employment levels:

$$AP_L = \frac{\text{Total output produced}}{\text{Total amount of labour employed}} = \frac{Q}{L}.$$

This function is also plotted in Figure 8.2. Referring to the table: The AP column indicates, for example, that when two units of labour are employed and forty units of output are produced, the average production level of each worker is 20 units ($= 40/2$). When three workers produce 70 units, their average production is 23.3 ($= 70/3$), and so forth. Like the MP function, this one also increases and subsequently decreases, reflecting exactly the same productivity forces that are at work on the MP curve.

> **Average product of labour** is the number of units of output produced per unit of labour at different levels of employment.

The AP and MP functions intersect at the point where the AP is at its peak. This is no accident, and has a simple explanation. Imagine a softball player who is batting .280 coming into today's game—she has been hitting her way onto base 28 percent of the time when batting, so far this season. This is her average product, AP.

In today's game, if she bats .500 (hits her way to base on half of her at-bats), then she will improve her average. Today's batting (*MP*) at .500 therefore pulls up the season's *AP*. Accordingly, whenever the *MP* exceeds the *AP*, the *AP* is pulled up. By the same reasoning, if her *MP* is less than the season average, her average will be pulled down. It follows that the two functions must intersect at the peak of the *AP* curve. To summarize:

> If the *MP* exceeds the *AP*, then the *AP* increases;
> If the *MP* is less than the *AP*, then the *AP* declines.

While the owner of BDS may understand his productivity relations, his ultimate goal is to make profit, and for this he must figure out how productivity translates into cost.

8.4 Costs in the short run

The cost structure for the production of snowboards at Black Diamond is illustrated in Table 8.2. Employees are skilled and are paid a weekly wage of $1,000. The cost of capital is $3,000 and it is fixed, which means that it does not vary with output. As in Table 8.1, the number of employees and the output are given in the first two columns. The following three columns define the capital costs, the labour costs, and the sum of these in producing different levels of output. We use the terms **fixed**, **variable**, and **total costs** to define the cost structure of a firm. Fixed costs do not vary with output, whereas variable costs do, and total costs are the sum of fixed and variable costs. To keep this example as simple as possible, we will ignore the cost of raw materials. We could add an additional column of costs, but doing so will not change the conclusions.

Table 8.2: Snowboard production costs

Workers	Output	Capital cost fixed	Labour cost variable	Total costs	Average fixed cost	Average variable cost	Average total cost	Marginal cost
0	0	3,000	0	3,000				
1	15	3,000	1,000	4,000	200.0	66.7	266.7	66.7
2	40	3,000	2,000	5,000	75.0	50.0	125.0	40.0
3	70	3,000	3,000	6,000	42.9	42.9	85.7	33.3
4	110	3,000	4,000	7,000	27.3	36.4	63.6	25.0
5	145	3,000	5,000	8,000	20.7	34.5	55.2	28.6
6	175	3,000	6,000	9,000	17.1	34.3	51.4	33.3
7	200	3,000	7,000	10,000	15.0	35.0	50.0	40.0
8	220	3,000	8,000	11,000	13.6	36.4	50.0	50.0
9	235	3,000	9,000	12,000	12.8	38.3	51.1	66.7
10	240	3,000	10,000	13,000	12.5	41.7	54.2	200.0

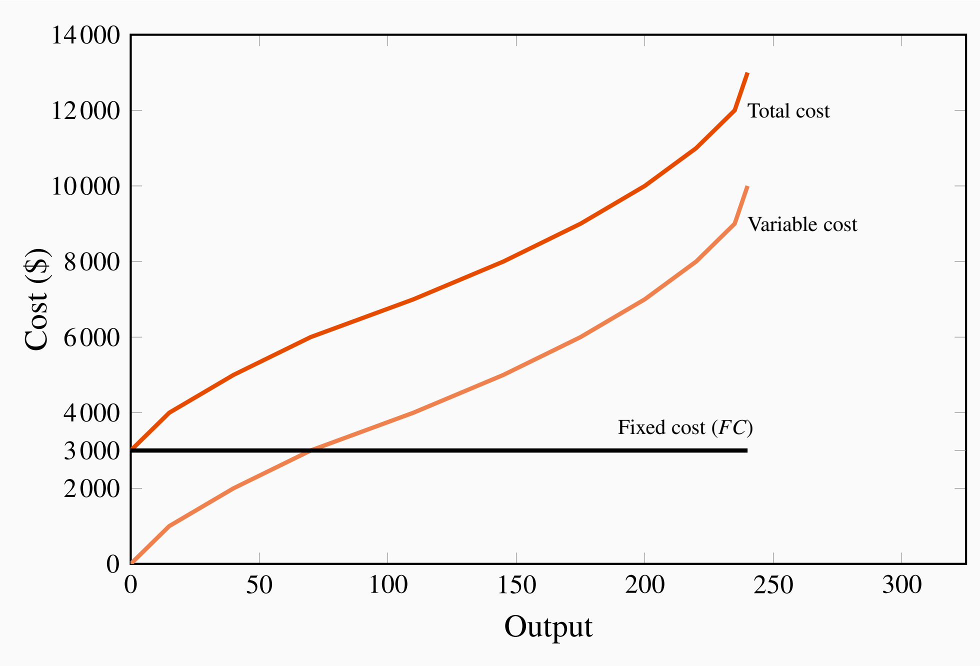
Fixed costs are costs that are independent of the level of output.

Variable costs are related to the output produced.

Total cost is the sum of fixed cost and variable cost.

Total costs are illustrated in Figure 8.3 as the vertical sum of variable and fixed costs. For example, Table 8.2 indicates that the total cost of producing 220 units of output is the sum of $3,000 in fixed costs plus $8,000 in variable costs. Therefore, at the output level 220 on the horizontal axis in Figure 8.3, the sum of the cost components yields a value of $11,000 that forms one point on the total cost curve. Performing a similar calculation for every possible output yields a series of points that together form the complete total cost curve.

Figure 8.3: Total cost curves



Total cost is the vertical sum of the variable and fixed costs.

Average costs are given in the next three columns of Table 8.2. Average cost is the cost per unit of output, and we can define an average cost corresponding to each of the fixed, variable, and total costs defined above. **Average fixed cost** (AFC) is the total fixed cost divided by output; **average variable cost** (AVC) is the total variable cost divided by output; and **average total cost** (ATC) is the total cost divided by output.

$$AFC = (\text{Fixed cost})/Q = FC/Q$$
$$AVC = (\text{Total variable costs})/Q = TVC/Q$$
$$ATC = AFC + AVC$$

Average fixed cost is the total fixed cost per unit of output.

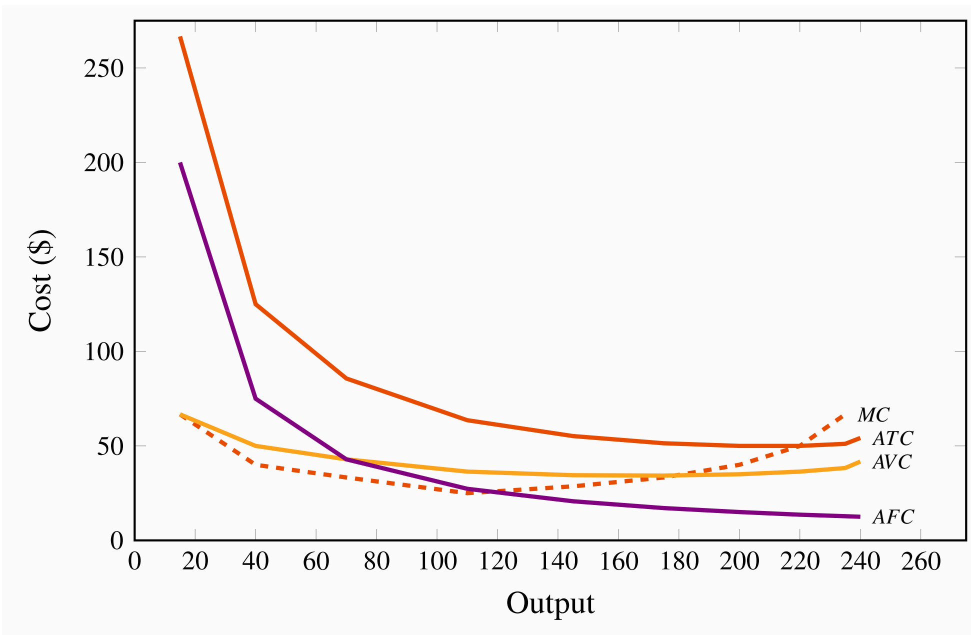
Average variable cost is the total variable cost per unit of output.

Average total cost is the sum of all costs per unit of output.

The productivity-cost relationship

Consider the **average variable cost - average product relationship**, as developed in column 7 of Table 8.2; its corresponding variable cost curve is plotted in Figure 8.4. In this example, AVC first decreases and then increases. The intuition behind its shape is straightforward (and realistic) if you have understood why productivity varies in the short run: The variable cost, which represents the cost of labour, is constant per unit of labour, because the wage paid to each worker does not change. However, each worker's productivity varies. Initially, when we hire more workers, they become more productive, perhaps because they have less 'down time' in switching between tasks. This means that the labour costs per snowboard must decline. At some point, however, the law of diminishing returns sets in: As before, each additional worker is paid a constant amount, but as productivity declines the labour cost per snowboard increases.

Figure 8.4: Average and marginal cost curves



The MC intersects the ATC and AVC at their minimum values. The AFC declines indefinitely as fixed costs are spread over a greater output.

In this numerical example the AP is at a maximum when six units of labour are employed and output is 175. This is also the point where the AVC is at a minimum. This maximum/minimum relationship is also illustrated in Figures 8.2 and 8.4.

Now consider the **marginal cost - marginal product relationship**. The **marginal cost** (MC) defines the cost of producing one more unit of output. In Table 8.2, the marginal cost of output is given in the final column. It is the additional cost of production divided by the additional number of units produced. For example, in going from 15 units of output to 40, total costs increase from \$4,000 to \$5,000. The MC is the cost of those additional units divided by the number of additional units. In this range of output, MC is $\$1,000/25 = \$40$. We could also calculate the MC as the addition to variable costs rather than the addition to total costs, because the *addition* to each is the same—fixed costs are fixed. Hence:

$$MC = \frac{\text{Change in total costs}}{\text{Change in output produced}} = \frac{\Delta TC}{\Delta Q}$$
$$= \frac{\text{Change in variable costs}}{\text{Change in output produced}} = \frac{\Delta TVC}{\Delta Q}.$$

Marginal cost of production is the cost of producing each additional unit of output.

Just as the behaviour of the AVC curve is determined by the AP curve, so too the behaviour of the MC is determined by the MP curve. When the MP of an additional worker exceeds the MP of the previous worker, this implies that the cost of the additional output produced by the last worker hired must be declining. To summarize:

> If the marginal product of labour increases, then the marginal cost of output declines;
> If the marginal product of labour declines, then the marginal cost of output increases.

In our example, the MP_L reaches a maximum when the fourth unit of labour is employed (or 110 units of output are produced), and this also is where the MC is at a minimum. This illustrates that the *marginal cost reaches a minimum at the output level where the marginal product reaches a maximum*.

The average total cost is the sum of the fixed cost per unit of output and the variable cost per unit of output. Typically, fixed costs are the dominant component of total costs at low output levels, but become less dominant at higher output levels. Unlike average variable costs, note that the average fixed cost must always decline with output, because a fixed cost is being spread over more units of output. Hence, when the ATC curve eventually increases, it is because the increasing variable cost component eventually dominates the declining AFC component. In our example, this occurs when output increases from 220 units (8 workers) to 235 (9 workers).

Finally, observe the interrelationship between the MC curve on the one hand and the ATC and AVC on the other. Note from Figure 8.4 that the MC cuts the AVC and the ATC at the minimum point of each of the latter. The logic behind this pattern is analogous to the logic of the relationship between marginal and average product curves: When the cost of an additional unit of output is less than the average, this reduces the average cost; whereas, if the cost of an additional unit of output is above the average, this raises the average cost. This must hold true regardless of whether we relate the MC to the ATC or the AVC.

> When the marginal cost is less than the average cost, the average cost must decline;
> When the marginal cost exceeds the average cost, the average cost must increase.

Notation: We use both the abbreviations ATC and AC to denote average total cost. The term 'average cost' is understood in economics to include both fixed and variable costs.

Teams and services

The choice faced by the producer in the example above is slightly 'stylized', yet it still provides an appropriate rule for analyzing hiring decisions. In practice, it is quite difficult to isolate or identify the marginal product of an individual worker. One reason is that individuals work in teams within organizations. The accounting department, the marketing department, the sales department, the assembly unit, the chief executive's unit are all composed of teams. Adding one more person to human resources may have no impact on the number of units of output produced by the company in a measurable way, but it may influence worker morale and hence longer-term productivity. Nonetheless, if we consider expanding, or contracting, any one department within an organization, management can attempt to estimate the net impact of additional hires (or layoffs) on the contribution of each team to the firm's profitability. Adding a person in marketing may increase sales, laying off a person in research and development may reduce costs by more than it reduces future value to the firm. In practice this is what firms do: they attempt to assess the contribution of each team in their organization to costs and revenues, and on that basis determine the appropriate number of employees.

The manufacturing sector of the macro economy is dominated, sizewise, by the services sector. But the logic that drives hiring decisions, as developed above, applies equally to services. For example, how does a law firm determine the optimal number of paralegals to employ per lawyer? How many nurses are required to support a surgeon? How many university professors are required to teach a given number of students?

All of these employment decisions involve optimization at the margin. The goal of the decision maker is not always profit, but she should attempt to estimate the cost and value of adding personnel at the margin.

8.5 Fixed costs and sunk costs

The distinction between fixed and variable costs is important for producers who are not making a profit. If a producer has committed himself to setting up a plant, then he has made a decision to incur a fixed cost. Having done this, he must now decide on a production strategy that will maximize profit. However, the price that consumers are willing to pay may not be sufficient to yield a profit. So, if Black Diamond Snowboards cannot make a profit, should it shut down? The answer is that if it can cover its variable costs, *having already incurred its fixed costs*, it should stay in production, at least temporarily. By covering the variable cost of its operation, Black Diamond is at least earning some return. A **sunk cost** is a fixed cost that has already been incurred and cannot be recovered. But if the pressures of the marketplace are so great that the total costs cannot be covered in the longer run, then this is not a profitable business and the firm should close its doors.

Is a fixed cost always a sunk cost? No: Any production that involves capital will incur a fixed cost component. Such capital can be financed in several ways however: It might be financed on a very short-term lease basis, or it might have been purchased by the entrepreneur. If it is leased on a month-to-month basis, an unprofitable entrepreneur who can only cover variable costs (and who does not foresee better market conditions ahead) can exit the industry quickly – by not renewing the lease on the capital. But an individual who has actually purchased equipment that cannot readily be resold has essentially sunk money into the fixed cost component of his production. This entrepreneur should continue to produce as long as he can cover variable costs.

> **Sunk cost** is a fixed cost that has already been incurred and cannot be recovered, even by producing a zero output.

R & D as a sunk cost

Sunk costs in the modern era are frequently in the form of research and development costs, not the cost of building a plant or purchasing machinery. The prototypical example is the pharmaceutical industry, where it is becoming progressively more challenging to make new drug breakthroughs – both because the 'easier' breakthroughs have already been made, and because it is necessary to meet tighter safety conditions attaching to new drugs. Research frequently leads to drugs that are not sufficiently effective in meeting their target. As a consequence, the pharmaceutical sector regularly writes off hundreds of millions of dollars of lost sunk costs – unfruitful research and development.

Finally, we need to keep in mind the opportunity costs of running the business. The owner pays himself a salary, and ultimately he must recognize that the survival of the business should not depend upon his drawing a salary that is less than his opportunity cost. As developed in Section 7.2, if he underpays himself in order to avoid shutting down, he might be better off in the long run to close the business and earn his opportunity cost elsewhere in the marketplace.

A dynamic setting

We need to ask why it might be possible to cover all costs in a longer run horizon, while in the near-term costs are not covered. The principal reason is that demand may grow, particularly for a new product. For example, in 2019 numerous cannabis producing firms were listed on the Canadian Securities Exchange, and collectively were valued at about fifty billion dollars. None had revenues that covered costs, yet investors poured money into this sector. Investors evidently envisaged that the market for legal cannabis would grow. As of 2020 it appears that these investors were excessively optimistic. Sales growth has been slow and stock valuations have plummeted.

8.6 Long-run production and costs

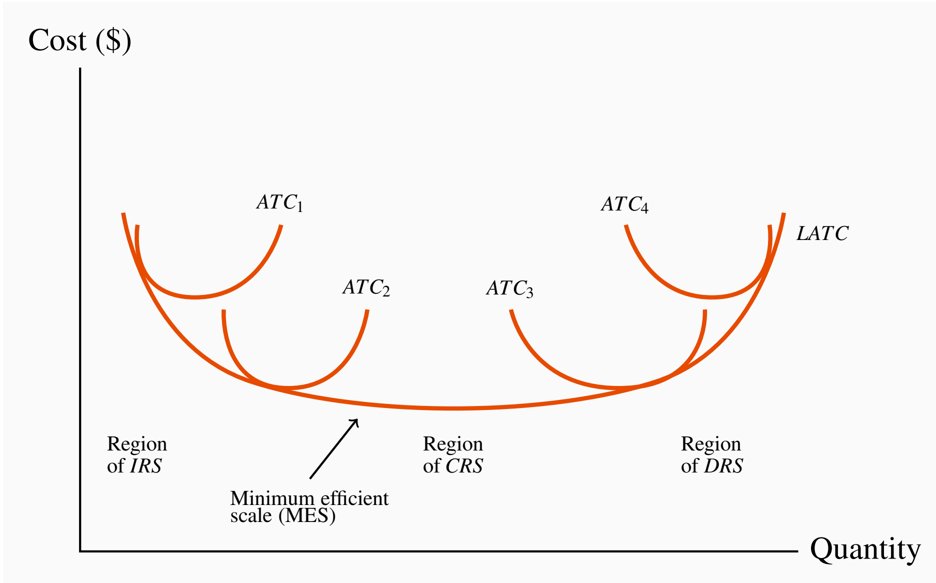
The snowboard manufacturer we portray produces a relatively low level of output; in reality, millions of snowboards are produced each year in the global market. Black Diamond Snowboards may have hoped to get a start by going after a local market—the "free-ride" teenagers at Mont Sainte Anne in Quebec or at Fernie in British Columbia. If this business takes off, the owner must increase production, take the business out of his garage and set up a larger-scale operation. But how will this affect his cost structure? Will he be able to produce boards at a lower cost than when he was producing a very limited number of boards each season? Real-world experience would indicate yes.

Production costs almost always decline when the *scale* of the operation initially increases. We refer to this phenomenon simply as **economies of scale**. There are several reasons why scale economies are encountered. One is that production flows can be organized in a more efficient manner when more is being produced. Another is that the opportunity to make greater use of task specialization presents itself; for example, Black Diamond Snowboards may be able to subdivide tasks within the laminating and packaging stations. With a larger operating scale the replacement of labor with capital may be economically efficient. If scale economies do define the real world, then a bigger plant—one that is geared to produce a higher level of output—should have an average total cost curve that is "lower" than the cost curve corresponding to the smaller scale of operation we considered in the example above.

Average costs in the long run

Figure 8.5 illustrates a possible relationship between the ATC curves for four different scales of operation. ATC_1 is the average total cost curve associated with a small-sized plant; think of it as the plant built in the entrepreneur's garage. ATC_2 is associated with a somewhat larger plant, perhaps one she has put together in a rented industrial or commercial space. The further a cost curve is located to the right of the diagram the larger the production facility it defines, given that output is measured on the horizontal axis. If there are economies associated with a larger scale of operation, then the average costs associated with producing larger outputs in a larger plant should be lower than the average costs associated with lower outputs in a smaller plant, assuming that the plants are producing the output levels they were designed to produce. For this reason, the cost curve ATC_2 and the cost curve ATC_3 each have a segment that is lower than the lowest segment on ATC_1. However, in Figure 8.5 the cost curve ATC_4 has moved upwards. What behaviours are implied here?

Figure 8.5: Long-run and short-run average costs



The long-run ATC curve, $LATC$, is the lower envelope of all short-run ATC curves. It defines the least cost per unit of output when all inputs are variable. Minimum efficient scale is that output level at which the $LATC$ is a minimum, indicating that further increases in the scale of production will not reduce unit costs.

In many production environments, beyond some large scale of operation, it becomes increasingly difficult to reap further cost reductions from specialization, organizational economies, or marketing economies. At such a point, the scale economies are effectively exhausted, and larger plant sizes no longer give rise to lower (short-run) ATC curves. This is reflected in the similarity of the ATC_2 and the ATC_3 curves. The pattern suggests that we have almost exhausted the possibilities of further scale advantages once we build a plant size corresponding to ATC_2. Consider next what is implied by the position of the ATC_4 curve relative to the ATC_2 and ATC_3 curves. The relatively higher position of the ATC_4 curve implies that unit costs will be higher in a yet larger plant. Stated differently: If we increase the scale of this firm to extremely high output levels, we are actually encountering **diseconomies of scale**. Diseconomies of scale imply that unit costs increase as a result of the firm's becoming too large: Perhaps co-ordination difficulties have set in at the very high output levels, or quality-control monitoring costs have risen. These coordination and management difficulties are reflected in increasing unit costs in the long run.

The terms **increasing**, **constant**, and **decreasing returns to scale** underlie the concepts of scale economies and diseconomies: Increasing returns to scale (IRS) implies that, when all inputs are increased by a given proportion, output increases more than proportionately. Constant returns to scale (CRS) implies that output increases in direct proportion to an equal proportionate increase in all inputs. Decreasing returns to scale (DRS) implies that an equal proportionate increase in all inputs leads to a less than proportionate increase in output.

Increasing returns to scale implies that, when all inputs are increased by a given proportion, output increases more than proportionately.

Constant returns to scale implies that output increases in direct proportion to an equal proportionate increase in all inputs.

Decreasing returns to scale implies that an equal proportionate increase in all inputs leads to a less than proportionate increase in output.

These are pure production function relationships, but, if the prices of inputs are fixed for producers, they translate directly into the various cost structures illustrated in Figure 8.5. For example, if a 40% increase in capital and labour use allows for better production flows than when in the smaller plant, and therefore yields more than a 40% increase in output, this implies that the cost per snowboard produced must fall in the new plant. In contrast, if a 40% increase in capital and labour leads to say just a 30% increase in output, then the cost per snowboard in the new larger plant must be higher. Between these extremes, there may be a range of relatively constant unit costs, corresponding to where the production relation is subject to constant returns to scale. In Figure 8.5, the falling unit costs output region has increasing returns to scale, the region that has relatively constant unit costs has constant returns to scale, and the increasing cost region has decreasing returns to scale.

Increasing returns to scale characterize businesses with large initial costs and relatively low costs of producing each unit of output. Computer chip manufacturers, pharmaceutical manufacturers, vehicle rental agencies, booking agencies such as *booking.com* or *hotels.com*, intermediaries such as *airbnb.com*, even brewers, all benefit from scale economies. In the beer market, brewing, bottling and shipping are all low-cost operations relative to the capital cost of setting up a brewery. Consequently, we observe surprisingly few breweries in any brewing company, even in large land-mass economies such as Canada or the US.

In addition to the four short-run average total cost curves, Figure 8.5 contains a curve that forms an envelope around the bottom of these short-run average cost curves. This envelope is the **long-run average total cost** ($LATC$) curve, because it defines average cost as we move from one plant size to another. Remember that in the long run both labour and capital are variable, and as we move from one short-run average cost curve to another, that is exactly what happens—all factors of production are variable. Hence, the collection of short-run cost curves in Figure 8.5 provides the ingredients for a long-run average total cost curve[1].

[1] Note that the long-run average total cost is not the collection of minimum points from each short-run average cost curve. The envelope of the short-run curves will pick up mainly points that are not at the minimum, as you will see if you try to draw the outcome. The intuition behind the definition is this: With increasing returns to scale, it may be better to build a plant size that operates with some spare capacity than to build one that is geared to producing a smaller output level. In building the larger plant, we can take greater advantage of the scale economies, and it may prove less costly to produce in such a plant than to produce with a smaller plant that has less unused capacity and does not exploit the underlying scale economies. Conversely, in the presence of decreasing returns to scale, it may be less costly to produce output in a plant that is used "overtime" than to use a larger plant that suffers from scale diseconomies.

$$LATC = (\text{Long-run total costs})/Q = LTC/Q$$

Long-run average total cost is the lower envelope of all the short-run ATC curves.

The particular range of output on the $LATC$ where it begins to flatten out is called the range of **minimum efficient scale**. This is an important concept in industrial policy, as we shall see in later chapters. At such an output level, the producer has expanded sufficiently to take advantage of virtually all the scale economies available.

Minimum efficient scale defines a threshold size of operation such that scale economies are almost exhausted.

In view of this discussion and the shape of the $LATC$ in Figure 8.5, it is obvious that economies of scale can also be defined in terms of the curvature of the $LATC$. Where the $LATC$ declines there are IRS, where the $LATC$ is flat there are CRS, where the $LATC$ slopes upward there are DRS.

Table 8.3: LATC elements for two plants (thousands $)

Q	AFC_1	$MC_1 = AVC_1$	ATC_1	AFC_2	$MC_2 = AVC_2$	ATC_2
20	50	30	**80**	100	25	125
40	25	30	**55**	50	25	75
60	16.67	30	**46.67**	33.33	25	58.33
80	12.5	30	**42.5**	25	25	50
100	10	30	**40**	20	25	45
120	8.33	30	**38.33**	16.67	25	41.67
140	7.14	30	**37.14**	14.29	25	39.29
160	6.25	30	**36.25**	12.5	25	37.5
180	5.56	30	**35.56**	11.11	25	36.11
200	*5*	*30*	**35**	*10*	*25*	**35**
220	4.55	30	34.55	9.09	25	**34.09**
240	4.17	30	34.17	8.33	25	**33.33**
260	3.85	30	33.85	7.69	25	**32.69**
280	3.57	30	33.57	7.14	25	**32.14**

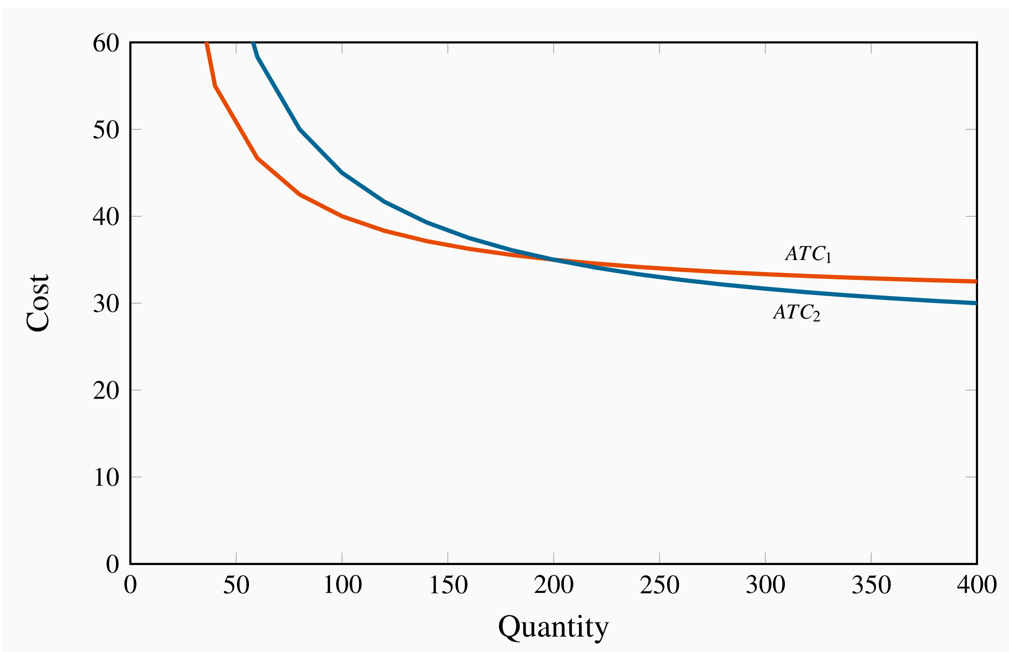
Plant 1 $FC = \$1$m. Plant 2 $FC = \$2$m. For $Q < 200, ATC_1 < ATC_2$; for $Q > 200, ATC_1 > ATC_2$; and for $Q = 200, ATC_1 = ATC_2$. $LATC$ defined by data in bold font.

Long-run costs – a simple numerical example

Kitt is an automobile designer specializing in the production of off-road vehicles sold to a small clientele. He has a choice of two (and only two) plant sizes; one involving mainly labour and the other employing robots extensively. The set-up (i.e. fixed) costs of these two assembly plants are \$1 million and \$2 million respectively. The advantage to having the more costly plant is that the pure production costs (variable costs) are less. The cost components are defined in Table 8.3. The variable cost (equal to the marginal cost here) is \$30,000 in the plant that relies primarily on labour, and \$25,000 in the plant that has robots. The ATC for each plant size is the sum of AFC and AVC. The AFC declines as the fixed cost is spread over more units produced. The variable cost per unit is constant in each case. By comparing the fourth and final columns, it is clear that the robot-intensive plant has lower costs if it produces a large number of vehicles. At an output of 200 vehicles the average costs in each plant are identical: The higher fixed costs associated with the robots are exactly offset by the lower variable costs at this output level.

The ATC curve corresponding to each plant size is given in Figure 8.6. There are two short-run ATC curves. The positions of these curves indicate that if the manufacturer believes he can produce at least 200 vehicles his unit costs will be less with the plant involving robots; but at output levels less than this his unit costs would be less in the labour-intensive plant.

Figure 8.6: LATC for two plants in \$000



The long-run average cost curve for this producer is the lower envelope of these two cost curves: ATC_1 up to output 200 and ATC_2 thereafter. Two features of this example are to be noted. First we do not encounter decreasing returns – the $LATC$ curve never increases. ATC_1 tends asymptotically to a lower bound of \$30, while ATC_2 tends towards \$25. Second, in the interests of simplicity we have assumed just two plant sizes are possible. With more possibilities on the introduction of robots we could imagine more short-run ATC curves which would form the lower-envelope $LATC$.

8.7 Technological change: globalization and localization

Technological change represents innovation that can reduce the cost of production or bring new products on line. As stated earlier, the very long run is a period that is sufficiently long for new technology to evolve and be implemented.

> **Technological change** represents innovation that can reduce the cost of production or bring new products on line.

Technological change has had an enormous impact on economic life for several centuries. It is not something that is defined in terms of the recent telecommunications revolution. The industrial revolution began in eighteenth century Britain. It was accompanied by a less well-recognized, but equally important, agricultural revolution. The improvement in cultivation technology, and ensuing higher yields, freed up enough labour to populate the factories that were the core of the industrial revolution[2]. The development and spread of mechanical power dominated the nineteenth century, and the mass production line of Henry Ford in autos or Andrew Carnegie in steel heralded in the twentieth century.

Globalization

The modern communications revolution has reduced costs, just like its predecessors. But it has also greatly sped up **globalization**, the increasing integration of national markets.

> **Globalization** is the tendency for international markets to be ever more integrated.

Globalization has several drivers: lower transportation and communication costs; reduced barriers to trade and capital mobility; the spread of new technologies that facilitate cost and quality control; different wage rates between developed and less developed economies. New technology and better communications have been critical in both increasing the minimum efficient scale of operation and reducing diseconomies of scale; they facilitate the efficient management of large companies.

The continued reduction in trade barriers in the post-World War II era has also meant that the effective marketplace has become the globe rather than the national economy for many products.

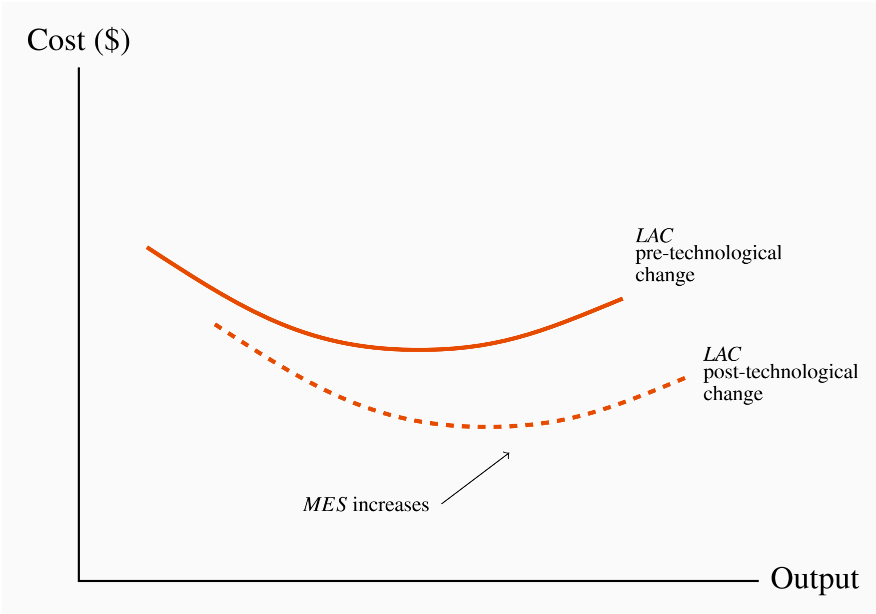
[2] Two English farmers changed the lives of millions in the 18th century through their technological genius. The first was Charles Townsend, who introduced the concept of crop rotation. He realized that continual use of soil for a single purpose drained the land of key nutrients. This led him ultimately to propose a five-year rotation that included tillage, vegetables, and sheep. This rotation reduced the time during which land would lie fallow, and therefore increased the productivity of land. A second game-changing technological innovation saw the introduction of the seed plough, a tool that facilitated seed sowing in rows, rather than scattering it randomly. This line sowing meant that weeds could be controlled more easily—weeds that would otherwise smother much of the crop. The genius here was a gentleman by the name of Jethro Tull.

Companies like *Apple*, *Microsoft*, and *Facebook* are visible worldwide. Globalization has been accompanied by the collapse of the Soviet Union, the adoption of an outward looking philosophy on the part of China, and an increasing role for the market place in India. These developments together have facilitated the outsourcing of much of the West's manufacturing to lower-wage economies.

But new technology not only helps existing companies grow large; it also enables new ones to start up. It is now cheaper for small producers to manage their inventories and maintain contact with their own suppliers.

The impact of technology is to reduce the cost of production, hence it will lower the average cost curve in both the short and long run. First, decreasing returns to scale become less probable due to improved communications, so the upward sloping section of the LRAC curve may disappear altogether. Second, capital costs are now lower than in earlier times, because much modern technology has transformed some fixed costs into variable cost: Both software and hardware functions can be subcontracted to specialty firms, who in turn may use cloud computing services, and it is thus no longer necessary to have a substantial in-house computing department. The use of almost-free software such as *Skype*, *Hangouts* and *WhatsApp* reduces communication costs. Advertising on social media is more effective and less costly than in traditional hard-print form. Hiring may be cheaper through *LinkedIn* than through a traditional human resources department. These developments may actually reduce the minimum efficient scale of operation because they reduce the need for large outlays on fixed capital. On the other hand, changes in technology may induce producers to use more capital and less labor, with the passage of time. That would increase the minimum efficient scale. An example of this phenomenon is in mining, or tunnel drilling, where capital investment per worker is greater than when technology was less developed. A further example is the introduction of robotic assistants in *Amazon* warehouses. In these scenarios the minimum efficient scale should increase, and that is illustrated in Figure 8.7.

Figure 8.7: Technological change and LAC



Technological change reduces the unit production cost for any output produced and may also increase the minimum efficient scale (MES) threshold.

The local diffusion of technology

The impacts of technological change are not just evident in a global context. Technological change impacts every sector of the domestic economy. For example, the modern era in dentistry sees specialists in root canals (endodentists) performing root canals in the space of a single hour with the help of new technology; dental implants into bone, as an alternative to dentures, are commonplace; crowns can be machined with little human intervention; and X-rays are now performed with about one hundredth of the power formerly required. These technologies spread and are adopted through several channels. Dental practices do not usually compete on the basis of price, but if they do not adopt best practices and new technologies, then community word of mouth will see patients shifting to more efficient operators.

Some technological developments are protected by patents. But patent protection rarely inhibits new and more efficient practices that in some way mimic patent breakthroughs.

8.8 Clusters, learning by doing, scope economies

Clusters

The phenomenon of a grouping of firms that specialize in producing related products is called a **cluster**. For example, Ottawa has more than its share of software development firms; Montreal has a disproportionate share of Canada's pharmaceutical producers and electronic game developers;

Calgary has its 'oil patch'; Hollywood has movies; Toronto is Canada's financial capital, San Francisco and Seattle are leaders in new electronic products. Provincial and state capitals have most of their province's bureaucracy. Clusters give rise to externalities, frequently in the form of ideas that flow between firms, which in turn result in cost reductions and new products.

> **Cluster**: a group of firms producing similar products, or engaged in similar research.

The most famous example of clustering is Silicon Valley, surrounding San Francisco, in California, the original high-tech cluster. The presence of a large group of firms with a common focus serves as a signal to workers with the right skill set that they are in demand in such a region. Furthermore, if these clusters are research oriented, as they frequently are, then knowledge spillovers benefit virtually all of the contiguous firms; when workers change employers, they bring their previously-learned skills with them; on social occasions, friends may chat about their work and interests and share ideas. This is a positive externality.

Learning by doing

Learning from production-related experiences frequently reduces costs: The accumulation of knowledge that is associated with having produced a large volume of output over a considerable time period enables managers to implement more efficient production methods and avoid errors. We give the term **learning by doing** to this accumulation of knowledge.

Examples abound, but the best known may be the continual improvement in the capacity of computer chips, whose efficiency has doubled about every eighteen months for several decades – a phenomenon known as Moore's Law. As *Intel Corporation* continues to produce chips it learns how to produce each succeeding generation of chips at lower cost. Past experience is key. Economies of scale and learning by doing therefore may not be independent: Large firms usually require time to grow or to attain a dominant role in their market, and this time and experience enables them to produce at lower cost. This lower cost in turn can solidify their market position further.

> **Learning by doing** can reduce costs. A longer history of production enables firms to accumulate knowledge and thereby implement more efficient production processes.

Economies of scope

Economies of scope define a production process if the production of multiple products results in lower unit costs per product than if those products were produced alone. Scope economies, therefore, define the returns or cost reductions associated with broadening a firm's product range.

Corporations like Proctor and Gamble do not produce a single product in their health line; rather, they produce first aid, dental care, and baby care products. Cable companies offer their customers TV, high-speed Internet, and telephone services either individually or packaged. A central component of some new-economy multi-product firms is a technology **platform** that can be used for multiple purposes. We shall analyze the operation of these firms in more detail in Chapter 11.

> **Economies of scope** occur if the unit cost of producing particular products is less when combined with the production of other products than when produced alone.

> **A platform** is a hardware-cum-software capital installation that has multiple production capabilities

Conclusion

Efficient production is critical to the survival of firms. Firms that do not adopt the most efficient production methods are likely to be left behind by their competitors. Efficiency translates into cost considerations, and the structure of costs in turn has a major impact on market type. Some sectors of the economy have very many firms (the restaurant business or the dry-cleaning business), whereas other sectors have few (internet providers or airlines). We will see in the following chapters how market structures depend critically upon the concept of scale economies that we have developed here.

Key Terms

Production function: a technological relationship that specifies how much output can be produced with specific amounts of inputs.

Technological efficiency means that the maximum output is produced with the given set of inputs.

Economic efficiency defines a production structure that produces output at least cost.

Short run: a period during which at least one factor of production is fixed. If capital is fixed, then more output is produced by using additional labour.

Long run: a period of time that is sufficient to enable all factors of production to be adjusted.

Very long run: a period sufficiently long for new technology to develop.

Total product is the relationship between total output produced and the number of workers employed, for a given amount of capital.

Marginal product of labour is the addition to output produced by each additional worker. It is also the slope of the total product curve.

Law of diminishing returns: when increments of a variable factor (labour) are added to a fixed amount of another factor (capital), the marginal product of the variable factor must eventually decline.

Average product of labour is the number of units of output produced per unit of labour at different levels of employment.

Fixed costs are costs that are independent of the level of output.

Variable costs are related to the output produced.

Total cost is the sum of fixed cost and variable cost.

Average fixed cost is the total fixed cost per unit of output.

Average variable cost is the total variable cost per unit of output.

Average total cost is the sum of all costs per unit of output.

Marginal cost of production is the cost of producing each additional unit of output.

Sunk cost is a fixed cost that has already been incurred and cannot be recovered, even by producing a zero output.

Increasing returns to scale implies that, when all inputs are increased by a given proportion, output increases more than proportionately.

Constant returns to scale implies that output increases in direct proportion to an equal proportionate increase in all inputs.

Decreasing returns to scale implies that an equal proportionate increase in all inputs leads to a less than proportionate increase in output.

Long-run average total cost is the lower envelope of all the short-run ATC curves.

Minimum efficient scale defines a threshold size of operation such that scale economies are almost exhausted.

Long-run marginal cost is the increment in cost associated with producing one more unit of output when all inputs are adjusted in a cost minimizing manner.

Technological change represents innovation that can reduce the cost of production or bring new products on line.

Globalization is the tendency for international markets to be ever more integrated.

Cluster: a group of firms producing similar products, or engaged in similar research.

Learning by doing can reduce costs. A longer history of production enables firms to accumulate knowledge and thereby implement more efficient production processes.

Economies of scope occur if the unit cost of producing particular products is less when combined with the production of other products than when produced alone.

A platform is a hardware-cum-software capital installation that has multiple production capabilities

EXERCISES FOR CHAPTER 8

Exercise 8.1 The relationship between output Q and the single variable input L is given by the form $Q = 5\sqrt{L}$. Capital is fixed. This relationship is given in the table below for a range of L values.

L	1	2	3	4	5	6	7	8	9	10	11	12
Q	5	7.07	8.66	10	11.18	12.25	13.23	14.14	15	15.81	16.58	17.32

(a) Add a row to this table and compute the MP.

(b) Draw the total product (TP) curve to scale, either on graph paper or in a spreadsheet.

(c) Inspect your graph to see if it displays diminishing MP.

Exercise 8.2 The TP for different output levels for Primitive Products is given in the table below.

Q	1	6	12	20	30	42	53	60	66	70
L	1	2	3	4	5	6	7	8	9	10

(a) Graph the TP curve to scale.

(b) Add a row to the table and enter the values of the MP of labour. Graph this in a separate diagram.

(c) Add a further row and compute the AP of labour. Add it to the graph containing the MP of labour.

(d) By inspecting the AP and MP graph, can you tell if you have drawn the curves correctly? How?

Exercise 8.3 A short-run relationship between output and total cost is given in the table below.

Output	0	1	2	3	4	5	6	7	8	9
Total Cost	12	27	40	51	61	70	80	91	104	120

(a) What is the total fixed cost of production in this example?

(b) Add four rows to the table and compute the TVC, AFC, AVC and ATC values for each level of output.

(c) Add one more row and compute the MC of producing additional output levels.

(d) Graph the MC and AC curves using the information you have developed.

Exercise 8.4 Consider the long-run total cost structure for the two firms A and B below.

Output	1	2	3	4	5	6	7
Total cost A	40	52	65	80	97	119	144
Total cost B	30	40	50	60	70	80	90

(a) Compute the long-run ATC curve for each firm.

(b) Plot these curves and examine the type of scale economies each firm experiences at different output levels.

Exercise 8.5 Use the data in Exercise 8.4,

(a) Calculate the long-run *MC* at each level of output for the two firms.

(b) Verify in a graph that these *LMC* values are consistent with the *LAC* values.

Exercise 8.6 *Optional*: Suppose you are told that a firm of interest has a long-run average total cost that is defined by the relationship $LATC = 4 + 48/q$.

(a) In a table, compute the $LATC$ for output values ranging from $1 \dots 24$. Plot the resulting $LATC$ curve.

(b) What kind of returns to scale does this firm never experience?

(c) By examining your graph, what will be the numerical value of the $LATC$ as output becomes very large?

(d) Can you guess what the form of the long-run *MC* curve is?

Part Four

Market Structures

Markets are all around us and they come in many forms. Some are on-line, others are physical. Some involve goods such as food and vehicles; others involve health provision or financial advice. Markets differ also by the degree of competition associated with each. For example the wholesale egg market is very homogeneous in that the product has minimal variation. The restaurant market offers food, but product variation is high. Each market has many suppliers.

In contrast to the egg and restaurant market, many other markets are characterized by just a few suppliers or in some cases just one. For example, passenger train services may have just a single supplier, and this supplier is therefore a monopolist. Pharmaceuticals tend to be supplied by a limited number of large international corporations plus a group of generic drug manufacturers. The internet and communications services market usually has only a handful of providers.

In this part we examine the reasons why markets take on different forms and display a variety of patterns of behaviour. We delve into the working of each market structure to understand why these markets retain their structure.

Chapter 9

Perfect competition

In this chapter we will explore:

9.1 The perfect competition paradigm

A competitive market is one that encompasses a very large number of suppliers, each producing a similar or identical product. Each supplier produces an output that forms a small part of the total market, and the sum of all of these individual outputs represents the production of that sector of the economy. Florists, barber shops, corner stores and dry cleaners all fit this description.

At the other extreme, a market that has just a single supplier is a monopolist. For example, the National Hockey League is the sole supplier of top-quality professional hockey games in North America; Hydro Quebec is a monopoly electricity distributor in Quebec; Via Rail is the only supplier of passenger rail services between Windsor, Ontario and the city of Quebec.

We use the word 'paradigm' in the title to this section: It implies that we will develop a *model* of supply behaviour for a market in which there are many small suppliers, producing essentially the same product, competing with one-another to meet the demands of consumers.

The structures that we call perfect competition and monopoly are extremes in the market place. Most sectors of the economy lie somewhere between these limiting cases. For example, the market for internet services usually contains several providers in any area – some provide using a fibre cable, others by satellite. The market for smart-phones in North America is dominated by two major players – *Apple* and *Samsung* (although there are several others). Hence, while these markets that have a limited number of suppliers are competitive in that they freely and perhaps fiercely compete for the buyer's expenditure, these are not **perfectly competitive** markets, because they do not have a very large number of suppliers.

In all of the models we develop in this chapter we will assume that the objective of firms is to maximize profit – the difference between revenues and costs.

> A **perfectly competitive** industry is one in which many suppliers, producing an identical product, face many buyers, and no one participant can influence the market.
>
> **Profit maximization** is the goal of competitive suppliers – they seek to maximize the difference between revenues and costs.

The presence of so many sellers in perfect competition means that each firm recognizes its own small size in relation to the total market, and that its actions have no perceptible impact on the market price for the good or service being traded. Each firm is therefore a *price taker*—in contrast to a monopolist, who is a *price setter*.

The same 'smallness' characteristic was assumed when we examined the demands of individuals earlier. Each buyer takes the price as given. He or she is not big enough to be able to influence the price. In contrast, when international airlines purchase or lease aircraft from *Boeing* or *Airbus*, they negotiate over the price and other conditions of supply. The market models underlying these types of transactions are examined in Chapter 11.

Hence, when we describe a market as being perfectly competitive we do not mean that other market types are not competitive; all market structure are competitive in the sense that the suppliers wish to make profit, and they produce as efficiently as possible in order to meet that goal.

9.2 Market characteristics

The key attributes of a perfectly competitive market are the following:

1. There must be *many firms*, each one so small that it cannot influence price or quantity in the industry, and powerless relative to the entire industry.
2. The *product must be standardized*. Barber shops offer a standard product, but a Lexus differs from a Ford. Barbers tend to be price takers, but Lexus does not charge the same price as Ford, and is a price setter.
3. Buyers are assumed to have *full information* about the product and its pricing. For example, buyers know that the products of different suppliers really are the same in quality.
4. There are *many buyers*.
5. There is *free entry and exit* of firms.

In terms of the demand curve that suppliers face, these market characteristics imply that the demand curve facing the perfectly competitive firm is horizontal, or infinitely elastic, as we defined in Chapter 4. In contrast, the demand curve facing the whole industry is downward sloping. The demand curve facing a firm is represented in Figure 9.1. It implies that the supplier can sell any output he chooses at the going price P_0. He is a small player in the market, and variations in his output have no perceptible impact in the marketplace. But what quantity should he choose, or what quantity will maximize his profit? The profit-maximizing choice is his target, and the *MC* curve plays a key role in this decision.

9.3 The firm's supply decision

The concept of **marginal revenue** is key to analyzing the supply decision of an individual firm. We have used marginal analysis at several points to date. In consumer theory, we saw how consumers balance the utility per dollar *at the margin* in allocating their budget. Marginal revenue is the additional revenue accruing to the firm from the sale of one more unit of output.

> **Marginal revenue** is the additional revenue accruing to the firm resulting from the sale of one more unit of output.

In perfect competition, a firm's marginal revenue (MR) is the price of the good. Since the price is constant for the individual supplier, each additional unit sold at the price P brings in the same additional revenue. Therefore, $P = MR$. For example, whether a dry cleaning business launders 10 shirts or 100 shirts per day, the price charged to customers is the same. This equality holds in no other market structure, as we shall see in the following chapters.

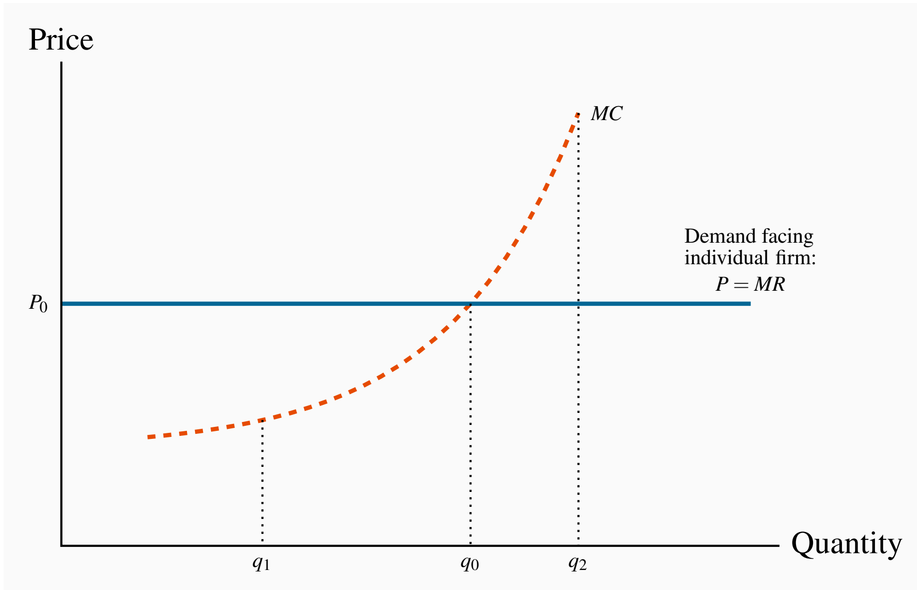
Supply in the short run

Recall how we defined the short run in the previous chapter: Each firm's plant size is fixed in the short run, so too is the number of firms in an industry. In the long run, each individual firm can change its scale of operation, and at the same time new firms can enter or existing firms can leave the industry.

Perfectly competitive suppliers face the choice of how much to produce at the going market price: That is, the amount that will maximize their profit. We abstract for the moment on how the price in the marketplace is determined. We shall see later in this chapter that it emerges as the value corresponding to the intersection of the supply and demand curves for the whole market – as described in Chapter 3.

The firm's *MC* curve is critical in defining the optimal amount to supply at any price. In Figure 9.1, *MC* is the firm's marginal cost curve in the short run. At the price P_0 the optimal amount to supply is q_0, the amount determined by the intersection of the *MC* and the demand. To see why, imagine that the producer chose to supply the quantity q_1. Such an output would leave the opportunity for further profit untapped. By producing one additional unit beyond q_1, the supplier would get P_0 in additional revenue and incur an additional cost that is less than P_0 in producing this unit. In fact, on every unit between q_1 and q_0 he can make a profit, because the *MR* exceeds the associated cost, *MC*. By the same argument, it makes no sense to increase output beyond q_0, to q_2 for example, because the cost of such additional units of output, *MC*, exceeds the revenue from them. *The MC therefore defines an optimal supply response*.

Figure 9.1: The competitive firm's optimal output



Here, q_0 represents the optimal supply decision when the price is P_0. At output q_1 the cost of additional units is less than the revenue from such units and therefore it is profitable to increase output beyond q_1. Conversely, at q_2 the MC of production exceeds the revenue obtained, and so output should be reduced.

Application Box 9.1: The law of one price

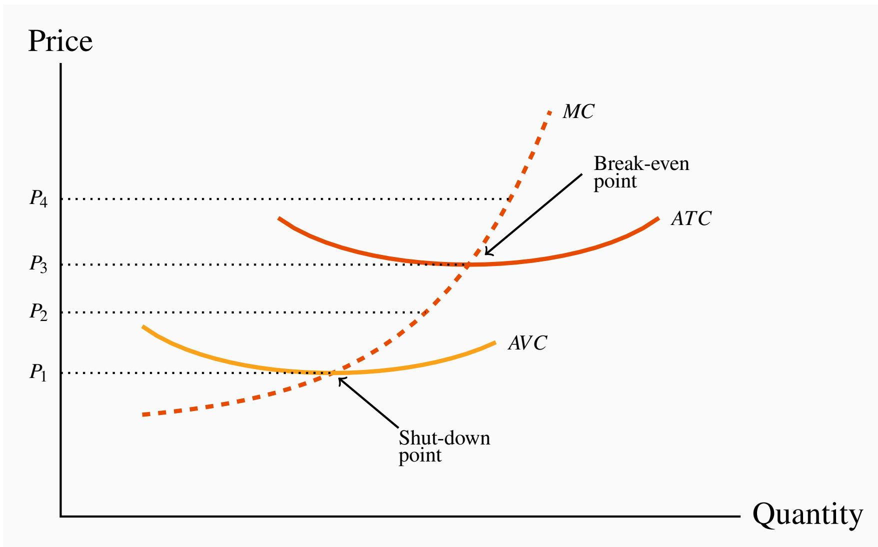
If information does not flow then prices in different parts of a market may differ and potential entrants may not know to enter a profitable market.

Consider the fishermen off the coast of Kerala, India in the late 1990s. Their market was studied by Robert Jensen, a development economist. Prior to 1997, fishermen tended to bring their fish to their home market or port. This was cheaper than venturing to other ports, particularly if there was no certainty regarding price. This practice resulted in prices that were high in some local markets and low in others – depending upon the daily catch. Frequently fish was thrown away in low-price markets even though it might have found a favourable price in another village's fish market.

This all changed with the advent of cell phones. Rather than head automatically to their home port, fishermen began to phone several different markets in the hope of finding a good price for their efforts. They began to form agreements with buyers before even bringing their catch to port. Economist Jensen observed a major decline in price variation between the markets that he surveyed. In effect the 'law of one price' came into being for sardines as a result of the introduction of cheap technology and the relatively free flow of information.

While the choice of the output q_0 is the best choice for the producer, Figure 9.1 does not tell us anything about profit. For that we need more information on costs. Accordingly, in Figure 9.2 the firm's AVC and ATC curves have been added to Figure 9.1. As explained in the previous chapter, the ATC curve includes both fixed and variable cost components, and the MC curve cuts the AVC and the ATC at their minima.

Figure 9.2: Short-run supply for the competitive firm



A price below P_1 does not cover variable costs, so the firm should shut down. Between prices P_1 and P_3, the producer can cover variable, but not total, costs and therefore should produce in the short run if fixed costs are 'sunk'. In the long run the firm must close if the price does not reach P_3. Profits are made if the price exceeds P_3. The short-run supply curve is the quantity supplied at each price. It is therefore the MC curve above P_1.

First, note that any price below P_3, which corresponds to the minimum of the ATC curve, yields no profit, since it does not enable the producer to cover all of his costs. This price is therefore called the **break-even price**. Second, any price below P_1, which corresponds to the minimum of the AVC, does not even enable the producer to cover variable costs. What about a price such as P_2, that lies between these? The answer is that, if the supplier has already incurred some fixed costs, he should continue to produce, provided he can cover his variable cost. But in the long run he must cover all of his costs, fixed and variable. Therefore, if the price falls below P_1, he should shut down, even in the short run. This price is therefore called the **shut-down price**. If a price at least equal to P_3 cannot be sustained in the long run, he should leave the industry. But at a price such as P_2 he can cover variable costs and therefore should continue to produce in the short run. His optimal output at P_2 is defined by the intersection of the P_2 line with the MC curve. The firm's **short-run supply curve** is, therefore, that portion of the MC curve above the minimum of the AVC.

To illustrate this more concretely, consider again the example of our snowboard producer, and imagine that he is producing in a perfectly competitive marketplace. How should he behave in response to different prices? Table 9.1 reproduces the data from Table 8.2.

Table 9.1: Profit maximization in the short run

Labour	Output	Total Revenue \$	Average Variable Cost	Average Total Cost \$	Marginal Cost \$	Total Cost \$	Profit
L	Q	TR	AVC	ATC	MC	TC	TR-TC
0	0					3,000	
1	15	1,050	66.67	266.67	66.67	4,000	−2,950
2	40	2,800	50.0	125.0	40.0	5,000	−2,200
3	70	4,900	42.86	85.71	33.33	6,000	−1,100
4	110	7,700	36.36	63.64	25.0	7,000	700
5	145	10,150	34.48	55.17	28.57	8,000	2,150
6	175	12,250	34.29	51.43	33.33	9,000	3,250
7	200	14,000	35.0	50.0	40.0	10,000	4,000
8	220	15,400	36.36	50.0	50.0	11,000	4,400
9	235	16,450	38.30	51.06	66.67	12,000	4,450
10	240	16,800	41.67	54.17	200.0	13,000	3,800

Output Price=\$70; Wage=\$1,000; Fixed Cost=\$3,000. The shut-down point occurs at a price of \$34.3, where the *AVC* attains a minimum. Hence no production, even in the short run, takes place unless the price exceeds this value. The break-even level of output occurs at a price of \$50, where the *ATC* attains a minimum.

The **shut-down price** corresponds to the minimum value of the *AVC* curve.

The **break-even price** corresponds to the minimum of the *ATC* curve.

The firm's **short-run supply curve** is that portion of the *MC* curve above the minimum of the *AVC*.

Suppose that the price is \$70. How many boards should he produce? The answer is defined by the behaviour of the *MC* curve. For any output less than or equal to 235, the *MC* is less than the price. For example, at $L = 9$ and $Q = 235$, the *MC* is \$66.67. At this output level, he makes a profit on the marginal unit produced, because the *MC* is less than the revenue he gets (\$70) from selling it.

But, at outputs above this, he registers a loss on the marginal units because the *MC* exceeds the revenue. For example, at $L = 10$ and $Q = 240$, the *MC* is \$200. Clearly, 235 snowboards is the optimum. To produce more would generate a loss on each additional unit, because the additional cost would exceed the additional revenue. Furthermore, to produce fewer snowboards would mean not availing of the potential for profit on additional boards.

His profit is based on the difference between revenue per unit and cost per unit at this output: $(P-ATC)$. Since the ATC for the 235 units produced by the nine workers is \$51.06, his profit margin is $\$70-\$51.06=\$18.94$ per board, and total profit is therefore $235\times\$18.94=\$4,450$.

Let us establish two other key outputs and prices for the producer. First, the shut-down point is the minimum of his AVC curve. Table 9.1 indicates that the price must be at least \$34.29 for him to be willing to supply *any* output, since that is the value of the AVC at its minimum. Second, the minimum of his ATC is at \$50. Accordingly, provided the price exceeds \$50, he will cover both variable and fixed costs and make a maximum profit when he chooses an output where $P=MC$, above $P=\$50$. It follows that the short-run supply curve for Black Diamond Snowboards is the segment of the MC curve in Figure 8.4 above the AVC curve.

Given that we have developed the individual firm's supply curve, the next task is to develop the industry supply curve.

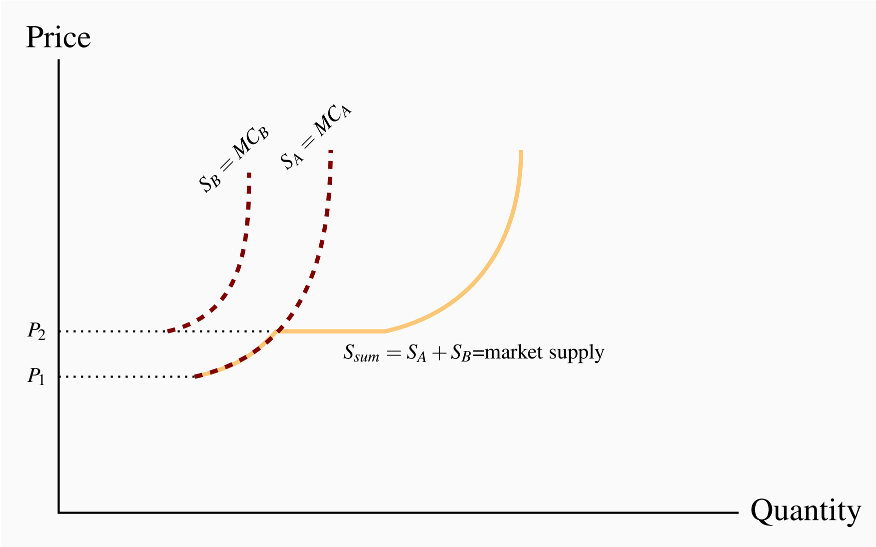
Industry supply in the short run

In Chapter 3 it was demonstrated that individual demands can be aggregated into an industry demand by summing them horizontally. The industry supply is obtained in exactly the same manner—by summing the firms' supply quantities across all firms in the industry.

To illustrate, imagine we have many firms, possibly operating at different scales of output and therefore having different short-run MC curves. The MC curves of two of these firms are illustrated in Figure 9.3. The MC of A is below the MC of B; therefore, B likely has a smaller scale of plant than A. Consider first the supply decisions in the price range P_1 to P_2. At any price between these limits, only firm A will supply output – firm B does not cover its AVC in this price range. Therefore, the joint contribution to industry supply of firms A and B is given by the MC curve of firm A. But once a price of P_2 is attained, firm B is now willing to supply. The S_{sum} schedule is the horizontal addition of their supply quantities. Adding the supplies of every firm in the industry in this way yields the **industry supply**.

> **Industry supply (short run)** in perfect competition is the horizontal sum of all firms' supply curves.

Figure 9.3: Deriving industry supply



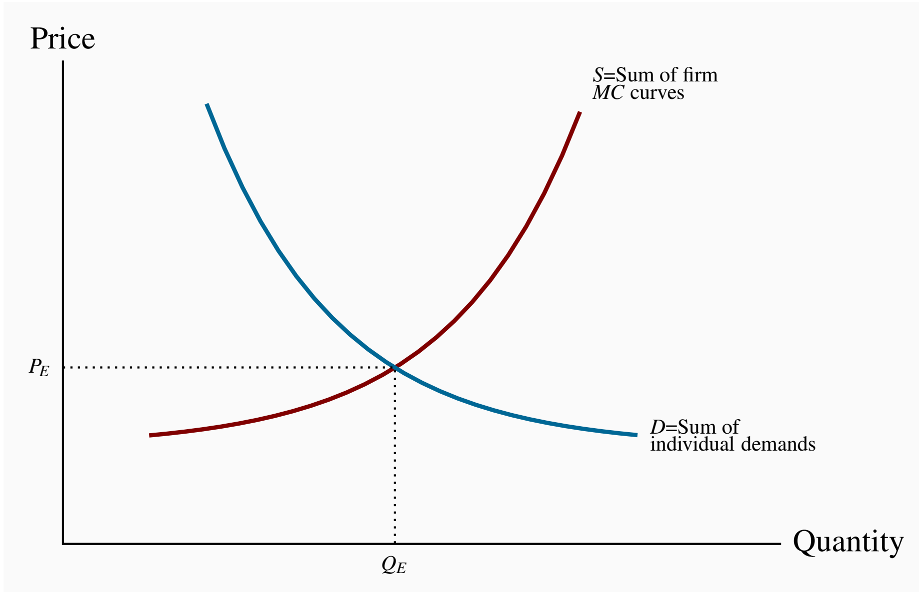
The marginal cost curves for firms A and B indicate that at any price below P_1 production is unprofitable and supply is therefore zero for both firms. At prices between P_1 and P_2 firm A is willing to supply, but not firm B. Consequently the market supply comes only from A. At prices above P_2 both firms are willing to supply. Therefore the market supply is the horizontal sum of each firm's supply.

Industry equilibrium

Consider next the industry equilibrium. Since the industry supply is the sum of the individual supplies, and the industry demand curve is the sum of individual demands, an equilibrium price and quantity (P_E, Q_E) are defined by the intersection of these industry-level curves, as in Figure 9.4. Here, each firm takes P_E as given (it is so small that it cannot influence the going price), and supplies an amount determined by the intersection of this price with its MC curve. The sum of such quantities is therefore Q_E.

> **Short-run equilibrium** in perfect competition occurs when each firm maximizes profit by producing a quantity where $P = MC$, provided the price exceeds the minimum of the average variable cost.

Figure 9.4: Market equilibrium



The market supply curve S is the sum of each firm's supply or MC curve above the shut-down price. D is the sum of individual demands. The market equilibrium price and quantity are defined by P_E and Q_E.

9.4 Dynamics: Entry and exit

We have now described the market and firm-level equilibrium in the short run. However, this equilibrium may be only temporary; whether it can be sustained or not depends upon whether profits (or losses) are being incurred, or whether all participant firms are making what are termed normal profits. Such profits are considered an essential part of a firm's operation. They reflect the opportunity cost of the resources used in production. Firms do not operate if they cannot make a minimal, or normal, profit level. Above such profits are **economic profits** (also called **supernormal profits**), and these are what entice entry into the industry.

Recall from Chapter 7 that accounting and economic profits are different. The economist includes opportunity costs in determining profit, whereas the accountant considers actual revenues and costs. In the example developed in Section 7.2 the entrepreneur recorded accounting profit, but not economic profit. Suppose now that the numbers were slightly different, and are as defined in Table 9.2: Felicity invests $250,000 in her business in the form of capital, as before. But she now has gross revenues of $165,000 and incurs a cost of $90,000 to buy the clothing wholesale that she then sells retail. She pays herself a salary of $35,000. If these numbers represent her balance sheet, then she records an accounting profit of $40,000.

Table 9.2: Economic profits

Sales		$165,000
Materials costs		$90,000
Wage costs		$35,000
Accounting profit		**$40,000**
Capital invested	$250,000	
Implicit return on capital at 4%		$10,000
Additional implicit wage costs		$20,000
Total implicit costs		**$30,000**
Economic profit		**$10,000**

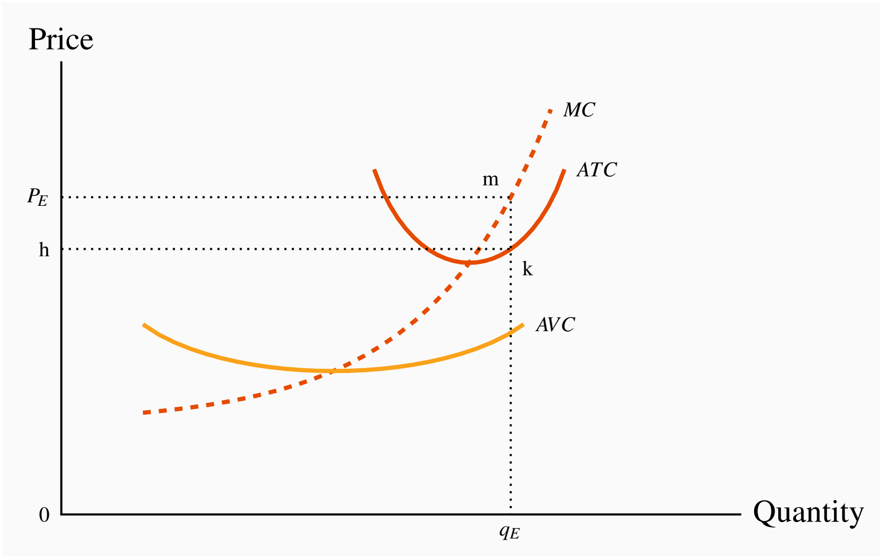
Her economic profit calculation must include opportunity costs. The opportunity cost of tying up $250,000 of capital, if the interest rate is 4%, amounts to $10,000. In addition, if Felicity could earn $55,000 in her best alternative job then an additional implicit cost of $20,000 must be considered. When these two opportunity (or implicit) costs are added to the balance sheet, her profit is reduced to $10,000. This is her economic profit. If Felicity's economic profit is representative of the retail clothing sector of the economy, then that profitability should attract new entrepreneurs. Our conclusion is that this sector of the economy *should experience new entrants and hence an outward shift of the supply curve*. In contrast, in the numerical example considered in Section 7.2, Felicity was experiencing losses (negative economic profits), and in the longer term she would have to consider leaving the business. If she and other suppliers exited, then the market supply curve would shift back to the left – representing a reduction in supply.

The critical point in this distinction between accounting and economic cost is that the decision to enter or leave a market in the longer term is based on what the entrepreneur can earn in the wider market place. That is, economic profits rather than accounting profits will determine the equilibrium number of firms in the long term. In terms of our cost curves, we will assume that the full economic costs are included in the various curves that we use. Consequently any profits (or losses) that arise are based upon the full economic costs of the firm's operation.

> **Economic (supernormal) profits** are those profits above normal profits that induce firms to enter an industry.

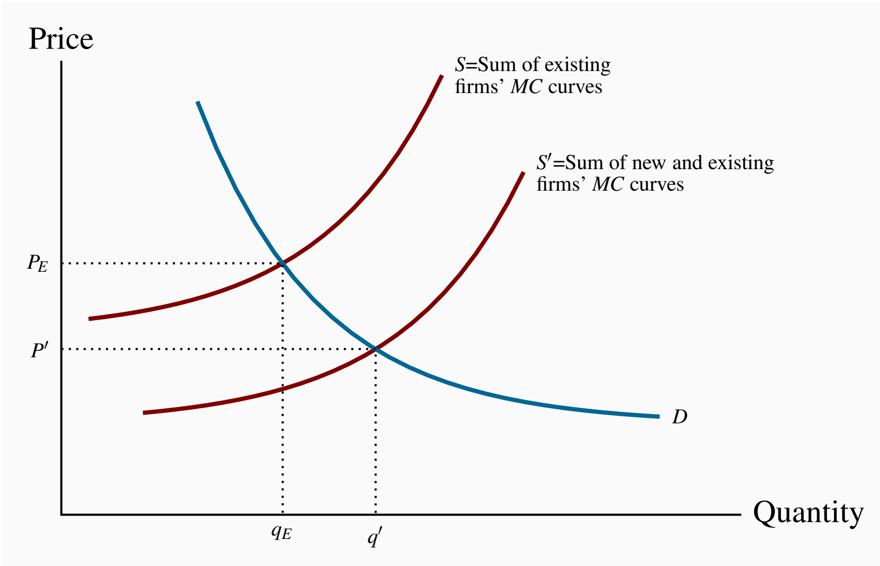
Let us return to our graphical analysis, and begin by supposing that the market equilibrium described in Figure 9.4 results in profits being made by some firms. Such an outcome is described in Figure 9.5, where the price exceeds the ATC. At the price P_E, a profit-making firm supplies the quantity q_E, as determined by its MC curve. On average, the cost of producing each unit of output, q_E, is defined by the point on the ATC at that output level, point k. Profit per unit is thus given by the value $(m-k)$ – the difference between revenue per unit and cost per unit. Total (economic) profit is therefore the area $P_E mkh$, which is quantity times profit per unit.

Figure 9.5: Short-run profits for the firm



At the price P_E, determined by the intersection of market demand and market supply, an individual firm produces the amount Q_E. The ATC of this output is k and therefore profit per unit is mk. Total profit is therefore $P_E$mkh$= 0q_E \times$mk$= TR-TC$.

Figure 9.6: Entry of firms due to economic profits



If economic profits result from the price P_E new firms enter the industry. This entry increases the market supply to S' and the equilibrium price falls to P'. Entry continues as long as economic profits are present. Eventually the price is driven to a level where only normal profits are made, and entry ceases.

While q_E represents an equilibrium for the firm, it is only a short-run, or temporary, equilibrium for the industry. The assumption of free entry and exit implies that the presence of economic profits will induce new entrepreneurs to enter and start producing. The impact of this dynamic is illustrated in Figure 9.6. An increased number of firms shifts supply rightwards to become S', thereby increasing the amount supplied at any price. The impact on price of this supply shift is evident: With an unchanged demand, the equilibrium price must fall.

How far will the price fall, and how many new firms will enter this profitable industry? As long as economic profits exist new firms will enter and the resulting increase in supply will continue to drive the price downwards. But, once the price has been driven down to the minimum of the ATC of a representative firm, there is no longer an incentive for new entrepreneurs to enter. Therefore, the **long-run industry equilibrium** is where the market price equals the minimum point of a firm's ATC curve. This generates normal profits, and there is no incentive for firms to enter or exit.

> A **long-run equilibrium** in a competitive industry requires a price equal to the minimum point of a firm's ATC. At this point, only normal profits exist, and there is no incentive for firms to enter or exit.

In developing this dynamic, we began with a situation in which economic profits were present. However, we could have equally started from a position of losses. With a market price between the minimum of the AVC and the minimum of the ATC in Figure 9.5, revenues per unit would exceed variable costs but not total costs per unit. When firms cannot cover their ATC in the long run, they will cease production. Such closures must reduce aggregate supply; consequently the market supply curve contracts, rather than expands as it did in Figure 9.6. The reduced supply drives up the price of the good. This process continues as long as firms are making losses. A final industry equilibrium is attained only when the price reaches a level where firms can make a normal profit. Again, this will be at the minimum of the typical firm's ATC.

Accordingly, the long-run equilibrium is the same, regardless of whether we begin from a position in which firms are incurring losses, or where they are making profits.

Application Box 9.2: Entry and exit: Oil rigs

Oil drilling is a competitive market. There are a large number of suppliers, information is ubiquitous, and entry and exit are relatively free.

In the years 2012 and 2013 the price of crude oil was around $100 US per barrel. Towards the end of 2014 the price of oil began to drop on world markets, and by early 2015 it fluctuated around $50. The response of drillers in the US was substantial and immediate. The number of active rigs declined dramatically. In terms of our economic model, certain suppliers exited; they moth-balled their rigs and waited for the price of oil to recover.

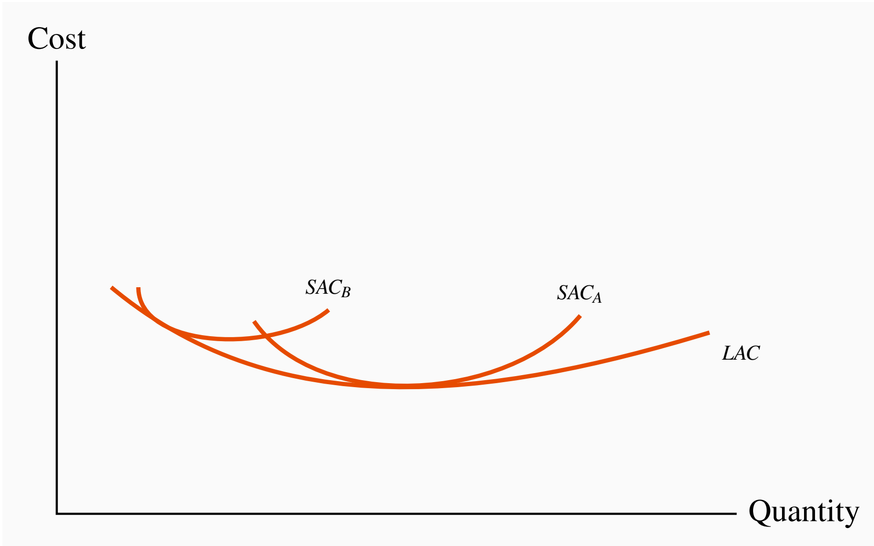
Another such cycle, even more pronounced, occurred in 2020. With the coronavirus pandemic, the demand for oil dropped and its price plummeted. Again, many firms shut down their rigs and had no choice but to sit out the price decline. A highly informative graphic is presented at https://tradingeconomics.com/united-states/crude-oil-rigs.

In addition to the decline in traditional oil recovery rigs, the number of operating shale crews declined by even greater amounts. Details at https://www.forbes.com/sites/davidblackmon/2020/05/12/a-grim-earnings-season-for-the-us-shale-business/#6f55a95a1cf2

9.5 Long-run industry supply

When aggregating the firm-level supply curves, as illustrated in Figure 9.3, we did not assume that all firms were identical. In that example, firm A has a cost structure with a lower AVC curve, since its supply curve starts at a lower dollar value. This indicates that firm A may have a larger plant size than firm B – one that puts A closer to the minimum efficient scale region of its long-run ATC curve.

Figure 9.7: Firms with different plant sizes



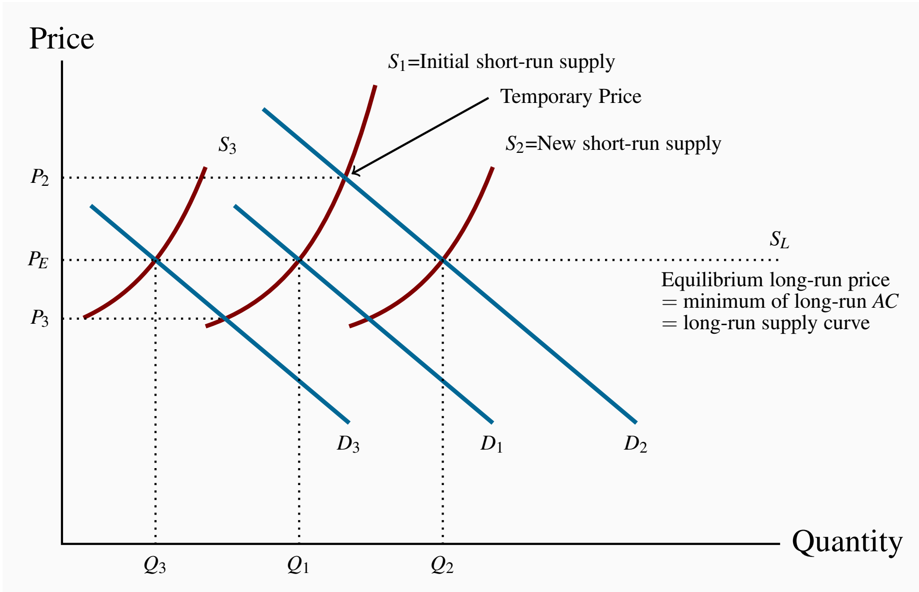
Firm B cannot compete with Firm A in the long run given that B has a less efficient plant size than firm A. The equilibrium long-run price equals the minimum of the *LAC*. At this price firm B must move to a more efficient plant size or make losses.

Can firm B survive with his current scale of operation in the long run? Our industry dynamics indicate that it cannot. The reason is that, provided *some* firms are making economic profits, new entrepreneurs will enter the industry and drive the price down to the minimum of the ATC curve *of those firms who are operating with the lowest cost plant size*. B-type firms will therefore be forced either to leave the industry or to adjust to the least-cost plant size—corresponding to the lowest point on its long-run ATC curve. Remember that the same technology is available to all firms; they each have the same long-run ATC curve, and may choose different scales of operation in the short run, as illustrated in Figure 9.7. But in the long run they must all produce using the minimum-cost plant size, or else they will be driven from the market.

This behaviour enables us to define a **long-run industry supply**. The long run involves the entry and exit of firms, and leads to a price corresponding to the minimum of the long-run ATC curve. Therefore, if the long-run equilibrium price corresponds to this minimum, *the long-run supply curve of the industry is defined by a particular price value—it is horizontal at the price corresponding to the minimum of the* $LATC$. More or less output is produced as a result of firms entering or leaving the industry, with those present always producing at the same unit cost in a long-run equilibrium.

Industry supply in the long run in perfect competition is horizontal at a price corresponding to the minimum of the representative firm's long-run ATC curve.

Figure 9.8: Long-run dynamics



The LR equilibrium price P_E is disturbed by a shift in demand from D_1 to D_2. With a fixed number of firms, P_2 results. Profits accrue at this price and entry occurs. Therefore the SR supply shifts outwards until these profits are eroded and the new equilibrium output is Q_2. If, instead, D falls to D_3 then firms exit because they make losses, S shifts back until the price is driven up sufficiently to restore normal profits. Different outputs are supplied in the long run at the same price P_E, therefore the long-run supply is horizontal at P_E.

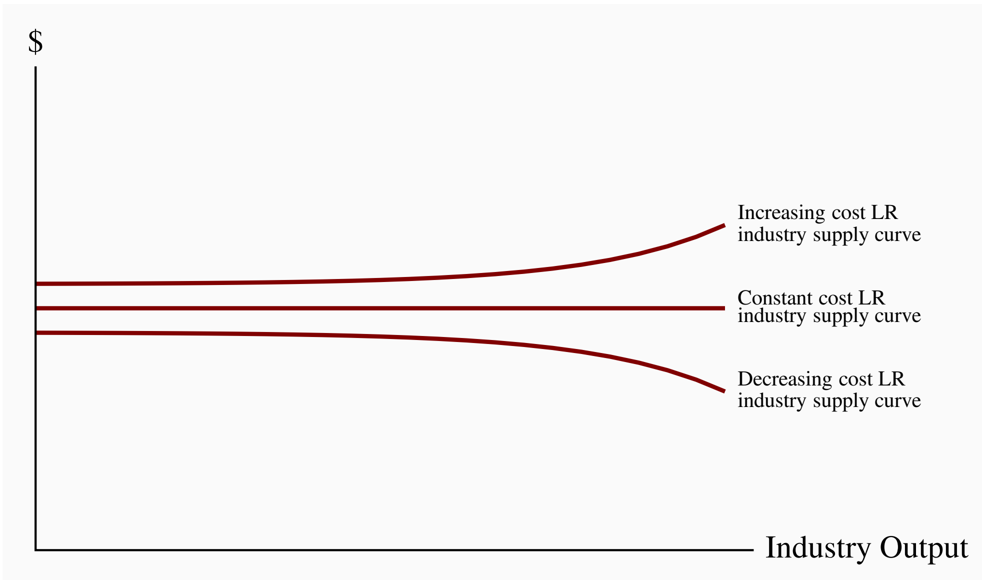
This industry's long-run supply curve, S_L, and a particular short-run supply are illustrated in Figure 9.8. Different points on S_L are attained when demand shifts. Suppose that, from an initial equilibrium Q_1, defined by the intersection of D_1 and S_1, demand increases from D_1 to D_2 because of a growth in income. With a fixed number of firms, the additional demand can be met only at a higher price (P_2), where each existing firm produces more using their existing plant size. The economic profits that result induce new operators to produce. This addition to the industry's production capacity shifts the short-run supply outwards and price declines until normal profits are once again being made. The new long-run equilibrium is at Q_2, with more firms each producing at the minimum of their long-run ATC curve, P_E.

The same dynamic would describe the industry reaction to a decline in demand—price would fall, some firms would exit, and the resulting contraction in supply would force the price back up to the long-run equilibrium level. This is illustrated by a decline in demand from D_1 to D_3.

Increasing and decreasing cost industries

While a horizontal long-run supply is the norm for perfect competition, in some industries costs increase with the scale of industry output; in others they decrease. This may be because all of the producers use a particular input that itself becomes more or less costly, depending upon the amount supplied.

Figure 9.9: Increasing and decreasing cost industries



When individual-supplier costs rise as the output of the industry increases we have an increasing cost supply curve for the industry in the long run. Conversely, when the costs of individual suppliers fall with the scale of the industry, we have a decreasing cost industry.

Decreasing cost sectors are those that benefit from a decline in the prices of their inputs as the size of their market expands. This is frequently because the suppliers of the inputs themselves can benefit from scale economies as a result of expansion in the market for the final good. A case in point has been the computer market, or the tablet market: As output in these markets has grown, the producers of videocards and random-access memory have benefited from scale economies and thus been able to sell these components at a lower price to the manufacturers of the final goods. An example of an **increasing cost** market is the market for landings and take-offs at airports. Airports are frequently limited in their ability to expand their size and build additional runways. In such markets, as use grows, planes about to land may have to adopt a circling holding pattern, while those departing encounter clearance delays. Such delays increase the time costs to passengers and the fuel and labour costs to the suppliers. Decreasing and increasing industry costs are reflected in the long-run industry supply curve by a downward-sloping segment or an upward sloping segment, as illustrated in Figure 9.9.

Increasing (decreasing) cost industry is one where costs rise (fall) for each firm because of the scale of industry operation.

9.6 Globalization and technological change

Globalization and technological change have had a profound impact on the way goods and services are produced and brought to market in the modern world. The cost structure of many firms has been reduced by *outsourcing to lower-wage economies*. Furthermore, the advent of the communications revolution has effectively *increased the minimum efficient scale for many industries*, as illustrated in Chapter 8 (Figure 8.7). Larger firms are less difficult to manage nowadays, and the *LAC* curve may not slope upwards until very high output levels are attained. The consequence is that some industries may not have sufficient "production space" to sustain a large number of firms. In order to reap the advantages of scale economies, firms become so large that they can supply a significant part of the market. They are no longer so small as to have no impact on the price.

Outsourcing and easier communications have in many cases simply eliminated many industries in the developed world. Garment making is an example. Some decades ago Quebec was Canada's main garment maker: Brokers dealt with 'cottage-type' garment assemblers outside Montreal and Quebec City. But ultimately the availability of cheaper labour in the developing world combined with efficient communications undercut the local manufacture. Most of Canada's garments are now imported. Other North American and European industries have been impacted in similar ways. Displaced labour has had to reskill, retool, reeducate itself, and either seek alternative employment in the manufacturing sector, or move to the service sector of the economy, or retire.

Globalization has had a third impact on the domestic economy, in so far as it *reduces the cost of components*. Even industries that continue to operate within national boundaries see a reduction in their cost structure on account of globalization's impact on input costs. This is particularly in evidence in the computing industry, where components are produced in numerous low-wage economies, imported to North America and assembled into computers domestically. Such components are termed *intermediate goods*.

9.7 Efficient resource allocation

Economists have a particular liking for competitive markets. The reason is not, as is frequently thought, that we love competitive battles; it really concerns resource allocation in the economy at large. In Chapter 5 we explained why markets are frequently an excellent vehicle for transporting the economy's resources to where they are most valued: A perfectly competitive marketplace in which there are no externalities results in resources being used up to the point where the demand and supply prices are equal. If demand is a measure of marginal benefit and supply is a measure of marginal cost, then a perfectly competitive market ensures that this condition will hold in equilibrium. *Perfect competition, therefore, results in resources being used efficiently*.

Our initial reaction to this perspective may be: If market equilibrium is such that the quantity supplied always equals the quantity demanded, is not every market efficient? The answer is no.

As we shall see in the next chapter on monopoly, the monopolist's supply decision does not reflect the marginal cost of resources used in production, and therefore does not result in an efficient allocation in the economy.

Key Terms

Perfect competition: an industry in which many suppliers, producing an identical product, face many buyers, and no one participant can influence the market.

Profit maximization is the goal of competitive suppliers – they seek to maximize the difference between revenues and costs.

Marginal revenue is the additional revenue accruing to the firm resulting from the sale of one more unit of output.

Shut-down price corresponds to the minimum value of the AVC curve.

Break-even price corresponds to the minimum of the ATC curve.

Short-run supply curve for perfect competitor: the portion of the MC curve above the minimum of the AVC.

Industry supply (short run) in perfect competition is the horizontal sum of all firms' supply curves.

Short-run equilibrium in perfect competition occurs when each firm maximizes profit by producing a quantity where $P = MC$.

Economic (supernormal) profits are those profits above normal profits that induce firms to enter an industry. Economic profits are based on the opportunity cost of the resources used in production.

Long-run equilibrium in a competitive industry requires a price equal to the minimum point of a firm's ATC. At this point, only normal profits exist, and there is no incentive for firms to enter or exit.

Industry supply in the long run in perfect competition is horizontal at a price corresponding to the minimum of the representative firm's long-run ATC curve.

Increasing (decreasing) cost industry is one where costs rise (fall) for each firm because of the scale of industry operation.

Exercises for Chapter 9

Exercise 9.1 Wendy's Window Cleaning is a small local operation. Wendy presently cleans the outside windows in her neighbours' houses for \$36 per house. She does ten houses per day. She is incurring total costs of \$420, and of this amount \$100 is fixed. The cost per house is constant.

(a) What is the marginal cost associated with cleaning the windows of one house – we know it is constant?

(b) At a price of \$36, what is her break-even level of output (number of houses)?

(c) If the fixed cost is 'sunk' and she cannot increase her output in the short run, should she shut down?

Exercise 9.2 A manufacturer of vacuum cleaners incurs a constant variable cost of production equal to \$80. She can sell the appliances to a wholesaler for \$130. Her annual fixed costs are \$200,000. How many vacuums must she sell in order to cover her total costs?

Exercise 9.3 For the vacuum cleaner producer in Exercise 9.2:

(a) Draw the MC curve.

(b) Next, draw her AFC and her AVC curves.

(c) Finally, draw her ATC curve.

(d) In order for this cost structure to be compatible with a perfectly competitive industry, what must happen to her MC curve at some output level?

Exercise 9.4 Consider the supply curves of two firms in a competitive industry: $P = q_A$ and $P = 2q_B$.

(a) On a diagram, draw these two supply curves, marking their intercepts and slopes numerically (remember that they are really MC curves).

(b) Now draw a supply curve that represents the combined supply of these two firms.

Exercise 9.5 Amanda's Apple Orchard Productions Limited produces 10,000 kilograms of apples per month. Her total production costs at this output level are \$8,000. Two of her many competitors have larger-scale operations and produce 12,000 and 15,000 kilos at total costs of \$9,500 and \$11,000 respectively. If this industry is competitive, on what segment of the *LAC* curve are these producers producing?

Exercise 9.6 Consider the data in the table below. TC is total cost, TR is total revenue, and Q is output.

Q	0	1	2	3	4	5	6	7	8	9	10
TC	10	18	24	31	39	48	58	69	82	100	120
TR	0	11	22	33	44	55	66	77	88	99	110

(a) Add some extra rows to the table and for each level of output calculate the MR, the MC and total profit.

(b) Next, compute AFC, AVC, and ATC for each output level, and draw these three cost curves on a diagram.

(c) What is the profit-maximizing output?

(d) How can you tell that this firm is in a competitive industry?

Exercise 9.7 *Optional*: The market demand and supply curves in a perfectly competitive industry are given by: $Q_d = 30,000 - 600P$ and $Q_s = 200P - 2000$.

(a) Draw these functions on a diagram, and calculate the equilibrium price of output in this industry.

(b) Now assume that an additional firm is considering entering. This firm has a short-run MC curve defined by $MC = 10 + 0.5q$, where q is the firm's output. If this firm enters the industry and it knows the equilibrium price in the industry, what output should it produce?

Exercise 9.8 *Optional*: Consider two firms in a perfectly competitive industry. They have the same MC curves and differ only in having higher and lower fixed costs. Suppose the ATC curves are of the form: $400/q + 10 + (1/4)q$ and $225/q + 10 + (1/4)q$. The MC for each is a straight line: $MC = 10 + (1/2)q$.

(a) In the first column of a spreadsheet enter quantity values of 1, 5, 10, 15, 20,..., 50. In the following columns compute the ATC curves for each quantity value.

(b) Compute the MC at each output in the next column, and plot all three curves.

(c) Compute the break-even price for each firm.

(d) Explain why both of these firms cannot continue to produce in the long run in a perfectly competitive market.

Chapter 10

Monopoly

In this chapter we will explore:

10.1 Monopolies

In analyzing perfect competition we emphasized the difference between the industry and the individual supplier. The individual supplier is an atomistic unit with no market power. In contrast, a monopolist has a great deal of market power, for the simple reason that a **monopolist** is the sole supplier of a particular product and so really *is* the industry. The word monopoly, comes from the Greek words *monos*, meaning one, and *polein* meaning to sell. When there is just a single seller, our analysis need not distinguish between the industry and the individual firm. They are the same on the supply side.

Furthermore, the distinction between long run and short run is blurred, because a monopoly that continues to survive as a monopoly obviously sees no entry or exit. This is not to say that monopolized sectors of the economy do not evolve, they do. Sometimes they die, sometimes they evolve in a different role. For example, when digital cameras entered the market place in the eighties the Polaroid Land camera (which printed film straight out of the camera) 'died' because the demand side of the market lost interest. The Blackberry 'smart' phone had a virtual monopoly on this product into the new millennium until Apple and Nokia entered the market.

> A **monopolist** is the sole supplier of an industry's output, and therefore the industry and the firm are one and the same.

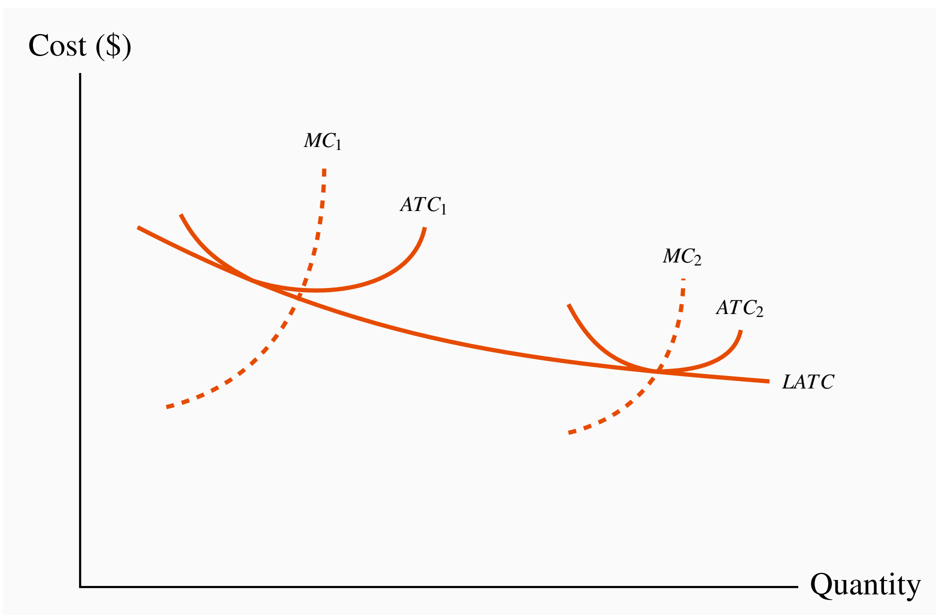
Monopolies can exist and exert their dominance in the market place for several reasons; scale economies, national policy, successful prevention of entry, research and development combined with patent protection.

Natural monopolies

Traditionally, monopolies were viewed as being 'natural' in some sectors of the economy. This means that scale economies define some industries' production and cost structures up to very high output levels, and that the whole market might be supplied at least cost by a single firm.

Consider the situation depicted in Figure 10.1. The long-run ATC curve declines indefinitely. There is no output level where average costs begin to increase. Imagine now having several firms, each producing with a plant size corresponding to the short-run average cost curve ATC_1, or alternatively a single larger firm using a plant size denoted by ATC_2. The small firms in this case cannot compete with the larger firm because the larger firm has lower production costs and can undercut the smaller firms, and *supply the complete market* in the process. Such a scenario is termed a **natural monopoly**.

Figure 10.1: A 'natural' monopolist



When LR average costs continue to decline at very high output, one large firm may be able to supply the industry at a lower unit cost than several smaller firms. With a plant size corresponding to ATC_2, a single supplier can supply the whole market, whereas several smaller firms, each with plan size corresponding to ATC_1, cannot compete with the larger firm on account of differential unit costs.

> **Natural monopoly**: one where the ATC of producing any output declines with the scale of operation.

Electricity distribution in some of Canada's provinces is in the hands of a single supplier – *Hydro Quebec* or *Hydro One* in Ontario, for example. These distributors are natural monopolies in the sense described above: Unit distribution costs decline with size. In contrast, electricity production is not 'naturally' a monopoly. Other suppliers were once thought of as 'natural' monopolies also, but are no longer. *Bell Canada* was considered to be a natural monopoly in the era of land lines: It would not make economic sense to run several sets of phone lines to every residence. But that was before the arrival of cell phones, broadband and satellites. *Canada Post* was also thought to be a natural monopoly, until the advent of *FEDEX*, *UPS* and other couriers proved otherwise. Invention can compete away a 'natural' monopoly.

In reality there are very few *pure* monopolies. *Facebook*, *Microsoft*, *Amazon*, *Apple*, *Netflix* and *Google* may be extraordinarily dominant in their markets, but they are not the only suppliers of the services or products that they offer. There exist other products that are similar.

National and Provincial Policy

Government policy can foster monopolies. Some governments are, or once were, proud to have a 'national carrier' in the airline industry – *Air Canada* in Canada or *British Airways* in the UK. The mail service was viewed as a symbol of nationhood in Canada and the US: *Canada Post* and the *US Postal* system are national emblems that have historic significance. They were vehicles for integrating the provinces or states at various points in the federal lives of these countries.

In the modern era, most of Canada's provinces have decided to create a provincial monopoly crown corporation for the sale of cannabis. But competition abounds in the form of an illegal market.

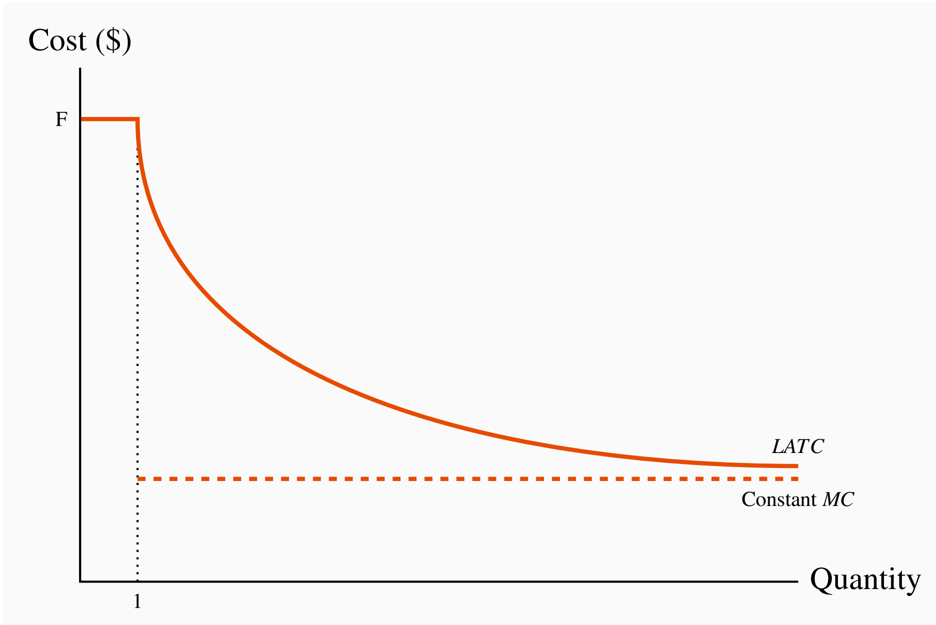
The down side of such nationalist policies is that they can be costly to the taxpayer. Industries that are not subject to competition can become fat and uncompetitive: Managers have insufficient incentives to curtail costs; unions realize the government is committed to sustain the monopoly and push for higher wages than under a more competitive structure, and innovation may be less likely to occur.

Maintaining barriers to entry

Monopolies can continue to survive if they are successful in preventing the entry of new firms and products. *Patents and copyrights* are one vehicle for preserving the sole-supplier role, and are certainly necessary to encourage firms to undertake the research and development (R&D) for new products.

Many corporations produce products that require a large up-front investment; this might be in the form of research and development, or the construction of costly production facilities. For example, *Boeing* or *Airbus* incurs billions of dollars in developing new aircraft; pharmaceuticals may have to invest a billion dollars to develop a new drug. However, once such an investment is complete, the cost of producing each unit of output may be relatively low. This is particularly true in pharmaceuticals. Such a phenomenon is displayed in Figure 10.2. In this case the average cost for a small number of units produced is high, but once the fixed cost is spread over an ever larger output, the average cost declines rapidly, and in the limit approaches the marginal cost. These production structures are common in today's global economy, and they give rise to markets characterized either by a single supplier or a small number of suppliers.

Figure 10.2: Fixed cost and constant marginal cost



With a fixed cost of producing the first unit of output equal to F and a constant marginal cost thereafter, the long-run average total cost, *LATC*, declines indefinitely and becomes asymptotic to the marginal cost curve.

This figure is useful in understanding the role of patents. Suppose that *Pharma A* spends one billion dollars in developing a new drug and has constant unit production costs thereafter, while *Pharma B* avoids research and development and simply imitates *Pharma A*'s product. Clearly *Pharma B* would have a *LATC* equal to its *LMC*, and would be able to undercut the initial developer of the drug. Such an outcome would discourage investment in new products and the economy at large would suffer as a consequence. Economies would be worse off if protection is not provided to the developers of new products because, if such protection is not offered, potential developers will not have the incentive to incur the up-front investment required.

While copyright and patent protection is legal, *predatory pricing* is an illegal form of entry barrier, and we explore it more fully in Chapter 14. An example would be where an existing firm that sells nationally may deliberately undercut the price of a small local entrant to the industry. Airlines with a national scope are frequently accused of posting low fares on flights in regional markets that a new carrier is trying to enter.

Political lobbying is another means of maintaining monopolistic power. For example, the *Canadian Wheat Board* had fought successfully for decades to prevent independent farmers from marketing wheat. This Board lost its monopoly status in August 2012, when the government of the day decided it was not beneficial to consumers or farmers in general. Numerous 'supply management' policies are in operation all across Canada. Agriculture is protected by production quotas. All maple syrup in Quebec must be marketed through a single monopoly supplier.

Critical networks also form a type of barrier, though not always a monopoly. *Microsoft's Office* package has an almost monopoly status in word processing and spreadsheet analysis for the reason that so many individuals and corporations use it. The fact that we know a business colleague will be able to edit our documents if written in *Word*, provides us with an incentive to use *Word*, even if we might prefer *Wordperfect* as a vehicle for composing documents. We develop the concept of strategic entry prevention further in Chapter 11.

10.2 Profit maximizing behaviour

We established in the previous chapter that, in deciding upon a profit-maximizing output, any firm should produce up to the point where the additional cost equals the additional revenue from a unit of output. What distinguishes the supply decision for a monopolist from the supply decision of the perfect competitor is that the monopolist faces a downward sloping demand. A monopolist is the sole supplier and therefore must meet the full market demand. This means that if more output is produced, the price must fall. We will illustrate the choice of a profit maximizing output using first a marginal-cost/marginal-revenue approach; then a supply/demand approach.

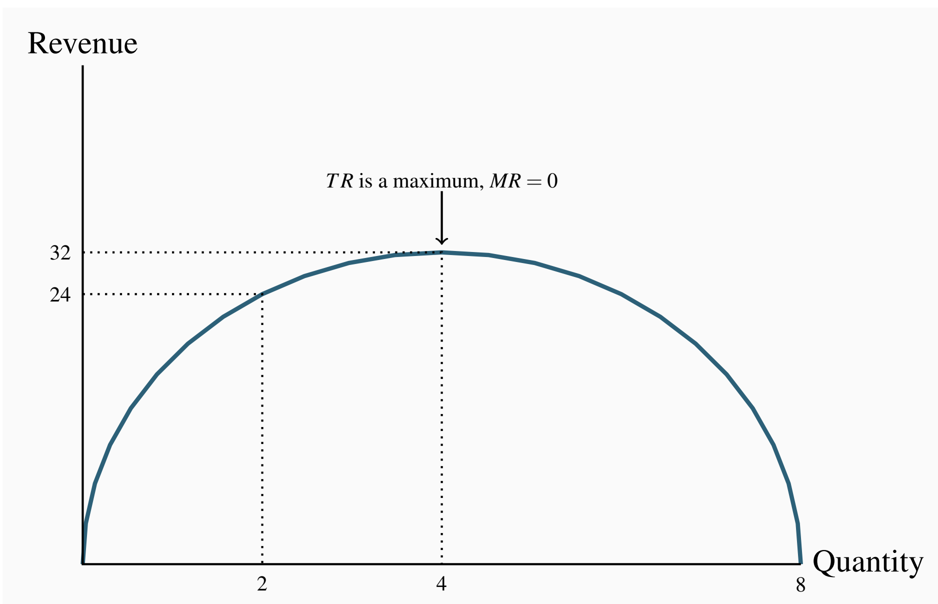
Marginal revenue and marginal cost

Table 10.1 displays price and quantity values for a demand curve in columns 1 and 2. Column 3 contains the sales revenue generated at each output. It is the product of price and quantity. Since the price denotes the revenue per unit, it is sometimes referred to as **average revenue**. The total revenue (TR) reaches a maximum at \$32, where 4 units of output are produced. A greater output necessitates a lower price on every unit sold, and in this case revenue falls if the fifth unit is brought to the market. Even though the fifth unit sells for a positive price, the price on the other 4 units is now lower and the net effect is to reduce total revenue. This pattern reflects what we examined in Chapter 4: As price is lowered from the highest possible value of \$14 (where 1 unit is demanded) and the corresponding quantity increases, revenue rises, peaks, and ultimately falls as output increases. In Chapter 4 we explained that this maximum revenue point occurs where the price elasticity is unity (-1), at the midpoint of a linear demand curve.

Table 10.1: A profit maximizing monopolist

Quantity (Q)	Price (P)	Total revenue (TR)	Marginal revenue (MR)	Marginal cost (MC)	Total cost (TC)	Profit
0	16					
1	14	14	14	2	2	12
2	12	24	10	3	5	19
3	10	30	6	4	9	21
4	8	32	2	5	14	18
5	6	30	-2	6	20	10
6	4	24	-6	7	27	-3
7	2	14	-10	8	35	-21

Figure 10.3: Total revenue and marginal revenue



When the quantity sold increases total revenue/expenditure initially increases also. At a certain point, further sales require a price that not only increases quantity, but reduces revenue on units already being sold to such a degree that TR declines – where the demand elasticity equals -1 (the mid point of a linear demand curve). Here the midpoint occurs at $Q = 4$. Where the TR is a maximum the $MR = 0$.

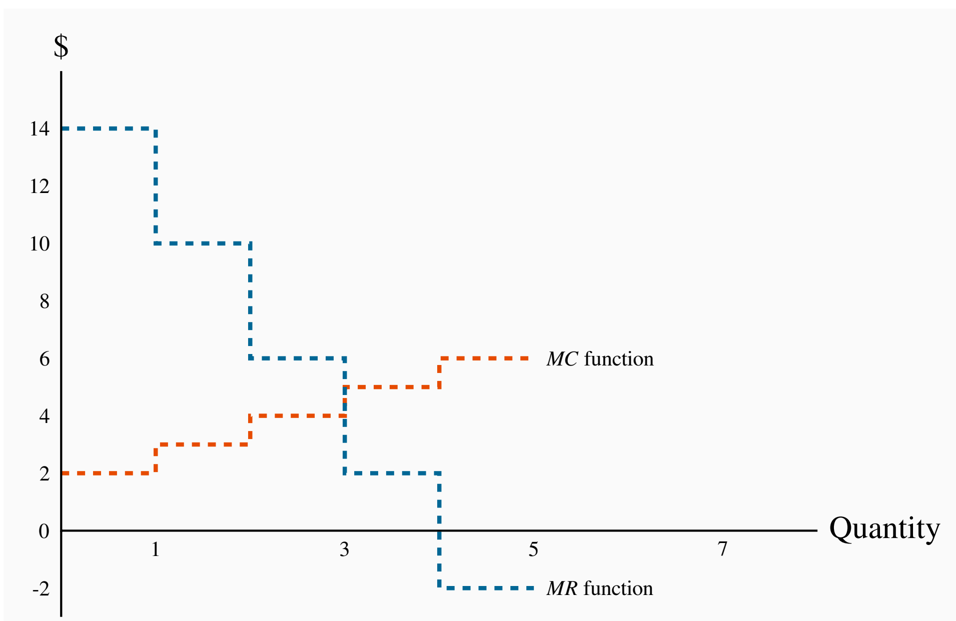
Related to the total revenue function is the **marginal revenue** function. It is the addition to total revenue due to the sale of one more unit of the commodity.

> **Marginal revenue** is the change in total revenue due to selling one more unit of the good.
>
> **Average revenue** is the price per unit sold.

The MR in this example is defined in the fourth column of Table 10.1. When the quantity sold increases from 1 unit to 2 units total revenue increases from \$14 to \$24. Therefore the marginal revenue associated with the second unit of output is \$10. When a third unit is sold TR increases to \$30 and therefore the MR of the third unit is \$6. As output increases the MR declines and eventually becomes negative – at the point where the TR is a maximum: If TR begins to decline then the additional revenue is by definition negative.

The MR function is plotted in Figure 10.4. It becomes negative when output increases from 4 to 5 units.

Figure 10.4: Monopolist's profit maximizing output



It is optimal for the monopolist to increase output as long as MR exceeds MC. In this case $MR > MC$ for units 1, 2 and 3. But for the fourth unit $MC > MR$ and therefore the monopolist would reduce total profit by producing it. He should produce only 3 units of output.

The optimal output

This producer has a marginal cost structure given in the fifth column of the table, and this too is plotted in Figure 10.4. Our profit maximizing rule from Chapter 8 states that it is optimal to produce a greater output as long as the additional revenue exceeds the additional cost of production on the next unit of output. In perfectly competitive markets the additional revenue is given by the fixed price for the individual producer, whereas for the monopolist the additional revenue is the marginal revenue. Consequently as long as MR exceeds MC for the next unit a greater output is profitable, but once MC exceeds MR the production of additional units should cease.

From Table 10.1 and Figure 10.4 it is clear that the optimal output is at 3 units. The third unit itself yields a profit of 2$, the difference between MR ($6) and MC ($4). A fourth unit however would reduce profit by $3, because the MR ($2) is less than the MC ($5). What price should the producer charge? The price, as always, is given by the demand function. At a quantity sold of 3 units, the corresponding price is $10, yielding total revenue of $30.

Profit is the difference between total revenue and total cost. In Chapter 8 we computed total cost as the average cost times the number of units produced. It can also be computed as the sum of costs associated with each unit produced: The first unit costs $2, the second $3 and the third $4. The total cost of producing 3 units is the sum of these dollar values: $\$9 = \$2 + \$3 + \$4$. The profit-maximizing output therefore yields a profit of $21 ($\$30 - \$9$).

Supply and demand

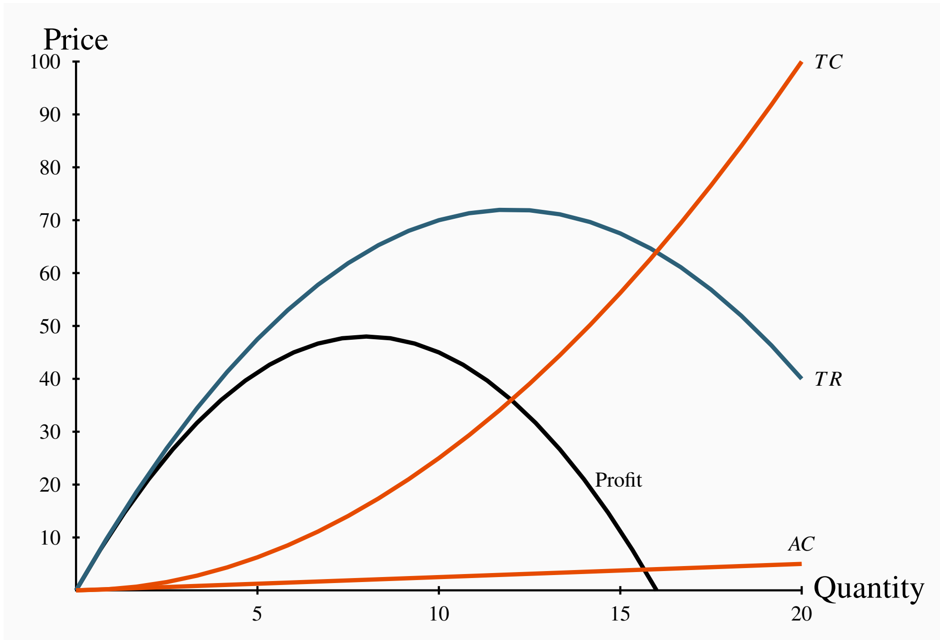
When illustrating market behaviour it is convenient to describe behaviour by simple linear supply and demand functions that are continuous, rather than the 'step' functions used in the preceding example. As explained in Chapter 5, in using continuous curves to represent a market we implicitly assume that a unit of output can be broken into subunits. In the example above we assumed that sales always involve one whole unit of the product being sold. In fact many goods can be sold in fractional units: Gasoline can be sold in fractions of a litre; fruits and vegetables can be sold in fractions of a kilogram, and so forth. Table 10.2 below furnishes the data for our analysis.

Table 10.2: Discrete quantities

Price	Quantity demanded	Total revenue	Total cost	Profit
12	0	0	0	0
11	2	22	1	21
10	4	40	4	36
9	6	54	9	45
8	8	64	16	48
7	10	70	25	45
6	12	72	36	36
5	14	70	49	21
4	16	64	64	0
3	18	54	81	-27
2	20	40	100	-60
1	22	22	121	-99
0	24	0	144	-144

The first two columns define the demand curve. Total revenue is the product of price and quantity and given in column 3. The cost data are given in column 4, and profit – the difference between total revenue and total cost is in the final column. Profit is maximized where the difference between revenue and cost is greatest; in this case where the output is 8 units. At lower or higher outputs profit is less. Figure 10.5 contains the curves defining total revenue (TR), total cost (TC) and profit. These functions can be obtained by mapping all of the revenue-quantity combinations, the cost-quantity combinations, and the profit-quantity combinations as a series of points, and joining these points to form the smooth functions displayed. The vertical axis is measured in dollars, the horizontal axis in units of output. Graphically, profit is maximized where the dollar difference between TR and TC is greatest; that is at the output where the vertical distance between the two curves is greatest. This difference, which is also defined by the profit curve, occurs at a value of 8 units, corresponding to the outcome in Table 10.2.

Figure 10.5: Total revenue, total cost & profit



At any quantity less than this output, profit would rise with additional output. This is because, from a less-than-optimal output, the additional revenue from increased sales exceeds the increased cost associated with producing those units: Stated differently, the marginal revenue would exceed the marginal cost. Conversely, outputs greater than the optimum result in a MR less than the associated MC. Accordingly, since outputs where $MR > MC$ are too low, and outputs where $MR < MC$ are too high, the optimum must be where the $MR = MC$. Hence, the equality between MR and MC is implied in this diagram at the output where the difference between TR and TC is greatest.

Note finally that total revenue is maximized where the TR curve reaches a peak. In this example that occurs at a value of 12 units of output. This is to be anticipated, as we learned in Chapter 4, because the midpoint of the demand schedule in Table 10.2 occurs at that value.

Figure 10.6: Market demand, the *MR* curve, and the monopolist's *AC* and *MC* curves

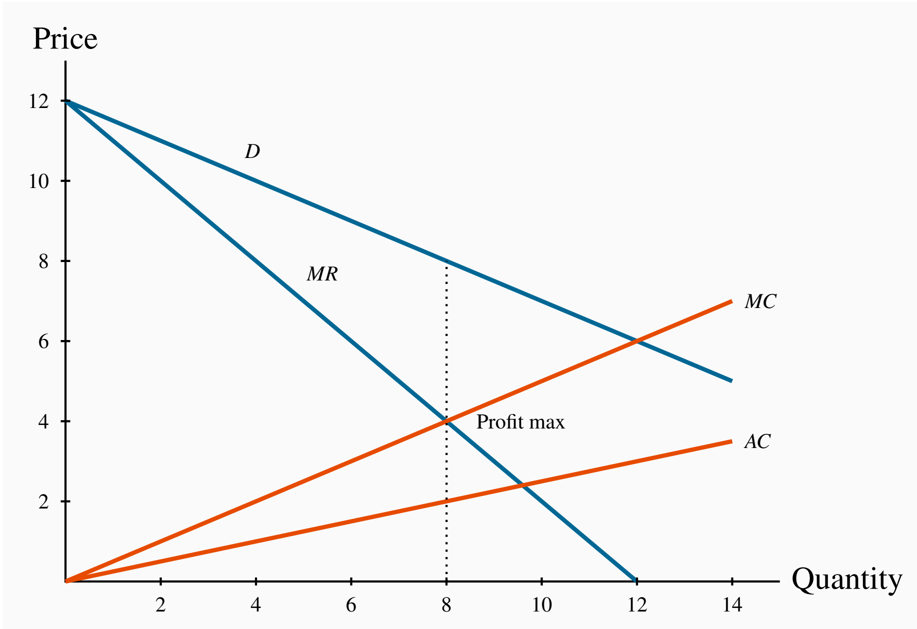


Figure 10.6 displays the demand curve for the market, the *MR* curve, and the monopolist's *MC* and *AC* curves. Consider first the marginal revenue curve. In contrast to the previous example, where only whole or integer units could be sold, in this example units can be sold in fractional amounts, and the *MR* curve must reflect this. To determine the position of the *MR* curve, note that with a straight-line demand curve total revenue is a maximum at the midpoint of the demand curve. Any increase in output results in reduced revenue: Stated differently, the marginal revenue becomes negative at that output. Up to that output the *MR* is positive, as illustrated in Figure 10.3. Accordingly, the *MR* curve must intersect the quantity axis midway between zero and the horizontal-axis intercept of the demand curve. Geometrically, since the *MR* intersects the quantity axis half way to the horizontal intercept of the demand curve, it must have a slope that is twice the slope of the demand curve.

By observing the data in columns 1 and 2 of the table, the demand curve intercepts are $\{\$12, 24\}$, and from above discussion the *MR* curve has intercepts $\{\$12, 12\}$. The *AC* is obtained by dividing *TC* by output in Table 10.2, and the *MC* can be also calculated as the change in total cost divided by the change in output from Table 10.2. The result of these calculations is displayed in Figure 10.6.

The profit maximizing output is 8 units, where $MC = MR$. The price at which 8 units can be sold is read from the demand curve[1], or the first column in Table 10.2. It is \$8. And, as expected, this price-quantity combination maximizes profit. Table 10.2 indicates that profit is maximized at \$48, at $q = 8$.

[1] It is not difficult to show that the demand curve corresponding to the data in the table is given by the expression $P = 12 - 0.5q$. Since the *MR* curve has twice the slope of the demand curve, it is given by $MR = 12 - q$. The data indicate that the *MC* curve can be written as $MC = 0.5q$ and the average cost curve by $AC = 0.25q$. Setting $MR = MC$ yields $q = 8$. At this output the demand curve implies the price is \$8. Profit is $(P - AC) \times q = (\$8 - \$2) \times 8 = \$48$.

Demand elasticity and marginal revenue

We have shown above that the MR curve cuts the horizontal axis at a quantity where the elasticity of demand is unity. We know from Chapter 4 that demand is elastic at points on the demand curve above this unit-elastic point. Furthermore, since the intersection of MR and MC must be at a positive dollar value (MC cannot be negative), then it must be the case that the *profit maximizing price for a monopolist always lies on the elastic segment of the demand curve*.

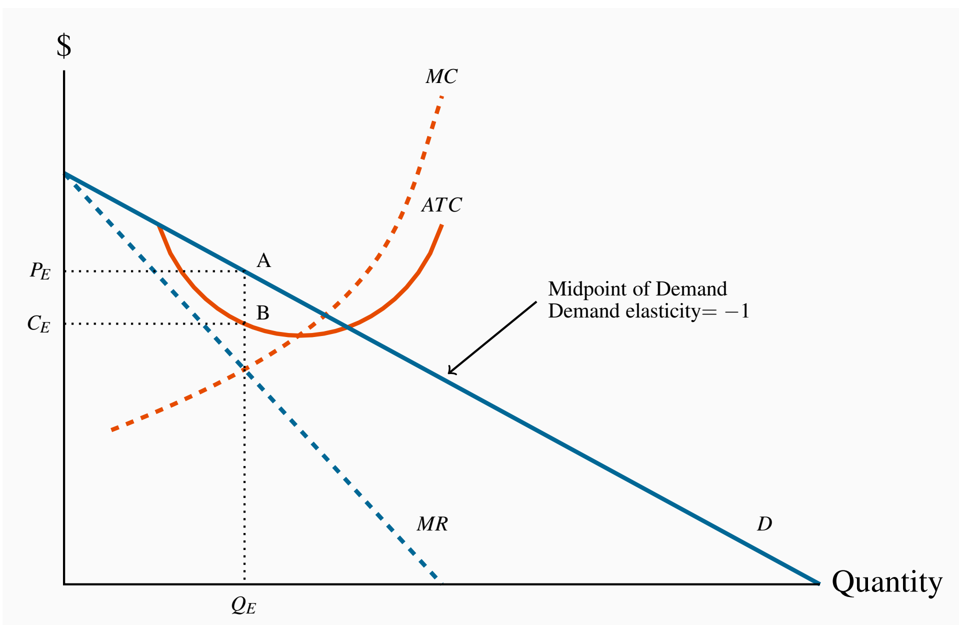
A general graphical representation

In Figure 10.7 we generalize the graphical representation of the monopoly profit maximizing output by allowing the MC and ATC curves to be nonlinear. The optimal output is at Q_E, where $MR = MC$, and the price P_E sustains that output. With the average cost known, profit per unit is AB, and therefore total profit is this margin multiplied by the number of units sold, Q_E.

Total profit is therefore $P_E ABC_E$.

Note that the monopolist may not always make a profit. Losses could result in Figure 10.7 if average costs were to rise so that the ATC were everywhere above the demand curve, or if the demand curve shifted down to being everywhere below the ATC curve. In the longer term the monopolist would have to either reduce costs or perhaps stimulate demand through advertising if she wanted to continue in operation.

Figure 10.7: The monopoly equilibrium



The profit maximizing output is Q_E, where $MC = MR$. This output can be sold at a price P_E. The cost per unit of Q_E is read from the ATC curve, and equals B. Per unit profit is therefore AB and total profit is P_EABC$_E$.

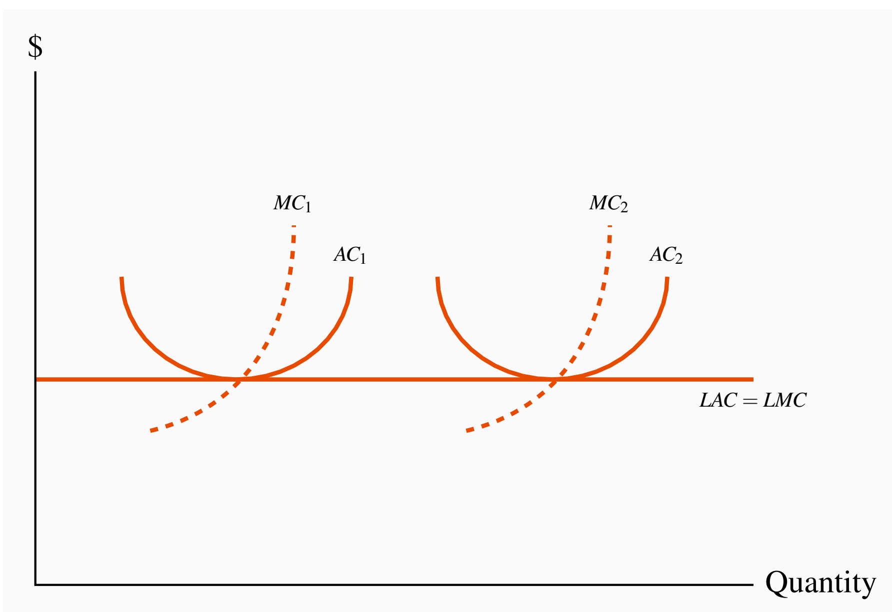
10.3 Long-run choices

Consider next the impact of a shift in demand upon the profit maximizing choice of this firm. A rightward shift in demand in Figure 10.7 also yields a new MR curve. The firm therefore chooses a new level of output, using the same profit maximizing rule: Set $MC = MR$. This output will be greater than the previous output, but again the price must be on an elastic portion of the new demand curve. If operating with the same plant size, the MC and ATC curves do not change and the new profit per unit is again read from the ATC curve.

By this stage the curious student will have asked: "What happens to plant size in the long run?" For example, is the monopolist in Figure 10.7 using the most appropriate plant size in the first place? Even if she is, should the monopolist consider adopting an expanded plant size in response to the shift in demand?

The answer is: In the long run the monopolist is free to choose whatever plant size is best. Her initial plant size might have been optimal for the demand she faced, but if it was, it is unlikely to be optimal for the larger scale of production associated with the demand shift. Accordingly, with the new demand curve, she must consider how much profit she could make using different plant sizes.

Figure 10.8: The monopolist's choice of plant size

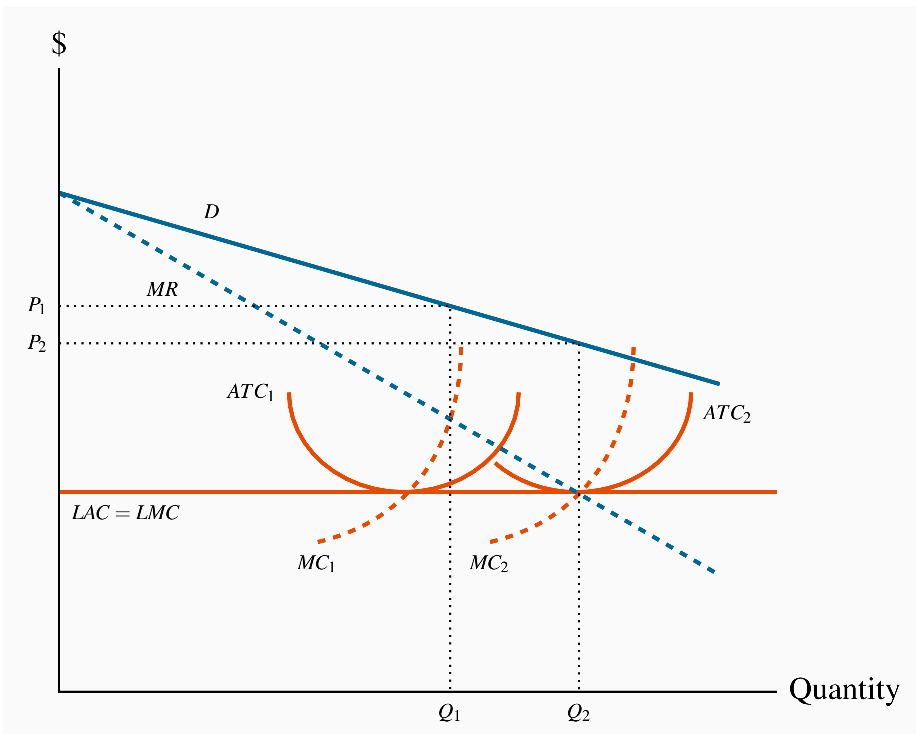


With constant returns to scale and constant prices per unit of labour and capital, a doubling of output involves exactly a doubling of costs. Thus, per unit costs, or average costs, are constant in the LR. Hence $LAC = LMC$, and each is constant.

To illustrate one possibility, we will think of this firm as having constant returns to scale at all output ranges, as displayed in Figure 10.8. (Our reasoning carries through if the LAC slopes downwards; the graph just becomes a little more complex.) The key characteristic of constant returns to scale is that a doubling of inputs leads to a doubling of output. Therefore, if the per-unit cost of inputs is fixed, a doubling of inputs (and therefore output) leads exactly to a doubling of costs. This implies that, when the firm varies its plant size *and* its labour use, the cost of producing each additional unit must be constant. The long-run marginal cost LMC is therefore constant and equals the ATC in the long run.

Figure 10.9 describes the market for this good. The optimal output and price are determined in the usual manner: Set $MC = MR$. If the monopolist has plant size corresponding to ATC_1, the optimal output is Q_1 and should be sold at the price P_1.The key issue now is: Given the demand conditions, could the monopolist make more profit by choosing a plant size that differs from the one corresponding to ATC_1?

Figure 10.9: Plant size in the long run



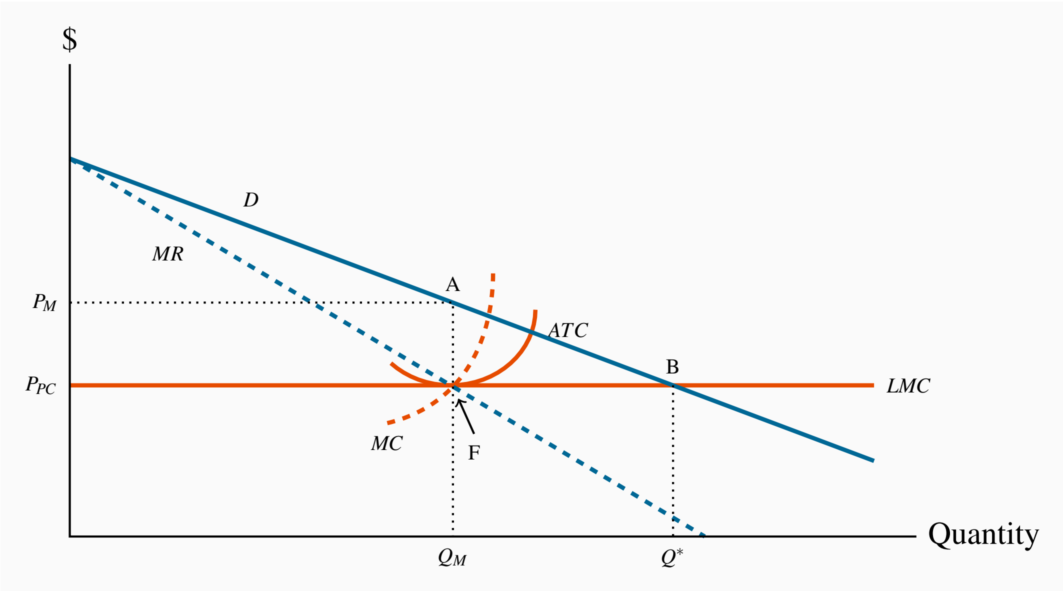
With demand conditions defined by D and MR, the optimal plant size is one corresponding to the point where $MR = MC$ in the long run. Therefore Q_2 is the optimal output and the optimal plant size corresponds to ATC_2. If the current plant is defined by ATC_1, then optimal SR production is Q_1.

In this instance the answer is a clear 'yes'. Her LMC curve is horizontal and so, by increasing output from Q_1 to Q_2 she earns a profit on each additional unit in that range, because the MR curve lies above the LMC curve. In order to produce the output level Q_2 at least cost she must choose a plant size corresponding to AC_2.

10.4 Output inefficiency

A characteristic of perfect competition is that it secures an efficient allocation of resources when there are no externalities in the market: Resources are used up to the point where their marginal cost equals their marginal value – as measured by the price that consumers are willing to pay. But a monopoly structure does not yield this output. Consider Figure 10.10.

Figure 10.10: Monopoly output inefficiency



A monopolist maximizes profit at Q_M. Here the value of marginal output exceeds cost. If output expands to Q^* a gain arises equal to the area ABF. This is the deadweight loss associated with the output Q_M rather than Q^*. If the monopolist's long-run *MC* is equivalent to a competitive industry's supply curve, then the deadweight loss is the cost of having a monopoly rather than a perfectly competitive market.

The monopolist's profit-maximizing output Q_M is where *MC* equals *MR*. This output is inefficient for the reason that we developed in Chapter 5: If output is increased beyond Q_M the additional benefit exceeds the additional cost of producing it. The additional benefit is measured by the willingness of buyers to pay – the market demand curve. The additional cost is the long-run *MC* curve under the assumption of constant returns to scale. Using the terminology from Chapter 5, there is a deadweight loss equal to the area ABF. This is termed **allocative inefficiency**.

> **Allocative inefficiency** arises when resources are not appropriately allocated and result in deadweight losses .

Perfect competition versus monopoly

The area ABF can also be considered as the efficiency loss associated with having a monopoly rather than a perfectly competitive market structure. In perfect competition the supply curve is horizontal. This is achieved by having firms enter and exit when more or less must be produced. Accordingly, *if the perfectly competitive industry's supply curve approximates the monopolist's long-run marginal cost curve*[2], we can say that if the monopoly were turned into a competitive industry, output would increase from Q_M to Q^*. The deadweight loss is one measure of the superiority of the perfectly competitive structure over the monopoly structure.

Note that this critique of monopoly is not initially focused upon profit. While monopoly profits are what frequently irk the public, we have focused upon resource allocation inefficiencies. But in a real sense the two are related: Monopoly inefficiencies arise through output being restricted, and it is this output reduction – achieved by maintaining a higher than competitive price – that gives rise to those profits. Nonetheless, there is more than just a shift in purchasing power from the buyer to the seller. Deadweight losses arise because output is at a level lower than the point where the MC equals the value placed on the good; thus the economy is sacrificing the possibility of creating additional surplus.

Given that monopoly has this undesirable inefficiency, what measures should be taken, if any, to counter the inefficiency? We will see what Canada's Competition Act has to say in Chapter 14 and also examine what other measures are available to control monopolies.

10.5 Price discrimination

A common characteristic in the pricing of many goods is that different individuals pay different prices for goods or services that are essentially the same. Examples abound: Seniors get a reduced rate for coffee in *Burger King*; hair salons charge women more than they charge men; bank charges are frequently waived for juniors. **Price discrimination** involves charging different prices to different consumers in order to increase profit.

> **Price discrimination** involves charging different prices to different consumers in order to increase profit.

A strict definition of discrimination involves different prices for *identical products*. We all know of a school friend who has been willing to take the midnight flight to make it home at school break at a price he can afford. In contrast, the business executive prefers the seven a.m. flight to arrive for a nine a.m. business meeting in the same city at several times the price. These are very mild forms of price discrimination, since a midnight flight (or a midday flight) is not a perfect substitute for an early morning flight. Price discrimination is practiced because buyers are willing to pay different amounts for a good or service, and the supplier may have a means of profiting from this. Consider the following example.

[2] We can think of such a transformation coming from a single supplier taking over a number of small suppliers, and the monopolist would thus be a multi-plant firm.

Family Flicks is the local movie theatre. It has two distinct groups of customers – those of prime age form one group; youth and seniors form the other. Family Flicks has done its market research and determined that each group accounts for 50 percent of the total market of 100 potential viewers per screening. It has also established that the prime-age group members are willing to pay \$12 to see a movie, while the seniors and youth are willing to pay just \$5. How should the tickets be priced?

Family Flicks has no variable costs, only fixed costs. It must pay a \$100 royalty to the movie maker each time it shows the current movie, and must pay a cashier and usher \$20 each. Total costs are therefore \$140, regardless of how many people show up – short-run *MC* is zero. On the pricing front, as illustrated in Table 10.3 below, if Family Flicks charges \$12 per ticket it will attract 50 viewers, generate \$600 in revenue and therefore make a profit of \$460.

Table 10.3: Price discrimination

	P=\$5	*P*=\$12	Twin price
No. of customers	100	50	
Total revenue	\$500	\$600	\$850
Total costs	\$140	\$140	\$140
Profit	\$360	\$460	\$710

In contrast, if it charges \$5 it can fill the theatre, because each of the prime-age individuals is willing to pay more than \$5, but the seniors and youth are now offered a price they too are willing to pay. However, the total revenue is now only \$500 ($100 \times \$5 = \$500$), and profits are reduced to \$360. It therefore decides to charge the high price and leave the theatre half-empty, because this strategy maximizes its profit.

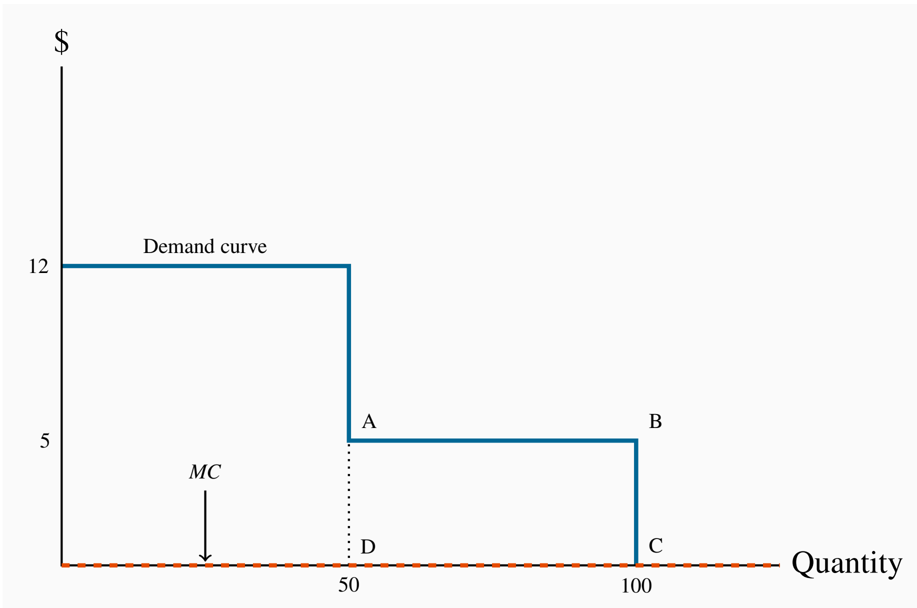
Suppose finally that the theatre is able to segregate its customers. It can ask the young and senior customers for identification upon entry, and in this way charge them a lower price, *while still maintaining the higher price to the prime-age customers*. If it can execute such a plan Family Flicks can now generate \$850 in revenue – \$600 from the prime-age group and \$250 from the youth and seniors groups. Profit soars to \$710.

There are two important conditions for this scheme to work:

1. The seller must be able to *segregate the market* at a reasonable cost. In the movie case this is achieved by asking for identification.
2. The second condition is that *resale must be impossible or impractical*. For example, we rule out the opportunity for young buyers to resell their tickets to the prime-age individuals. Sellers have many ways of achieving this – they can require immediate entry to the movie theatre upon ticket purchase, they can stamp the customer's hand, they can demand the showing of ID with the ticket when entering the theatre area.

Frequently we think of sellers who offer price reductions to specific groups as being generous. For example, hotels may levy only a nominal fee for the presence of a child, once the parents have paid a suitable rate for the room or suite in which a family stays. The hotel knows that if it charges too much for the child, it may lose the whole family as a paying unit. The coffee shop offering cheap coffee to seniors is interested in getting a price that will cover its variable cost and so contribute to its profit. It is unlikely to be motivated by philanthropy, or to be concerned with the financial circumstances of seniors.

Figure 10.11: Price discrimination at the movies



At $P = 12$, 50 prime-age individuals demand movie tickets. At $P = 5$, 50 more seniors and youths demand tickets. Since the *MC* is zero the efficient output is where the demand curve takes a zero value – where all 100 customers purchase tickets. Thus, any scheme that results in all 100 individuals buying ticket is efficient. Efficient output is at point C.

Price discrimination has a further interesting feature that is illustrated in Figure 10.11: It frequently *reduces the deadweight loss* associated with a monopoly seller!

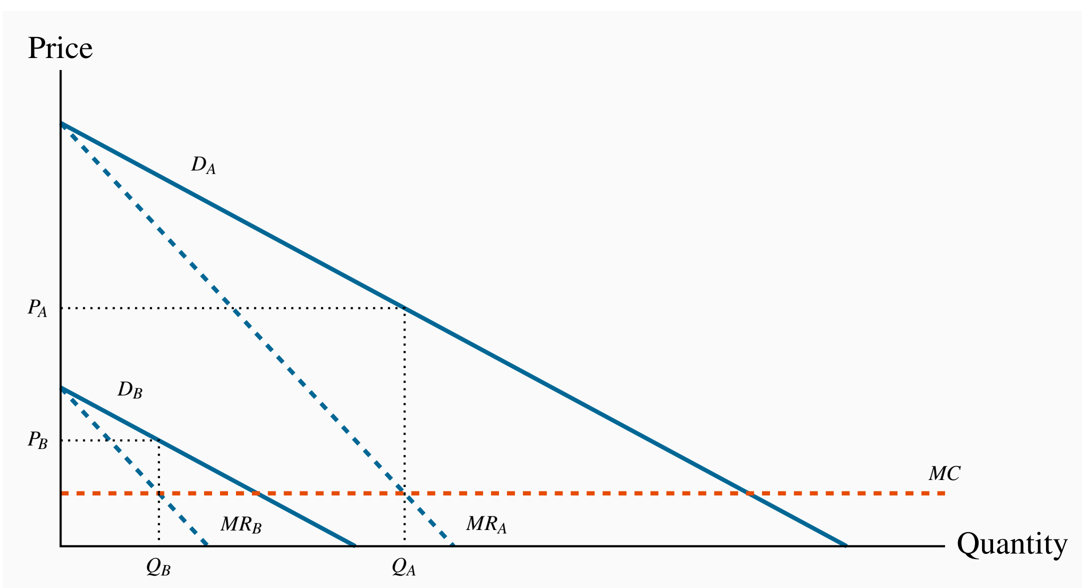
In our Family Flicks example, the profit maximizing monopolist that did not, or could not, price discriminate *left 50 customers unsupplied who were willing to pay $5 for a good that had a zero MC*. This is a deadweight loss of $250 because 50 seniors and youth valued a commodity at $5 that had a zero *MC*. Their demand was not met because, in the absence of an ability to discriminate between consumer groups, Family Flicks made more profit by satisfying the demand of the prime-age group alone. But in this example, by segregating its customers, the firm's profit maximization behaviour resulted in the DWL being eliminated, because it supplied the product to those additional 50 individuals. In this instance *price discrimination improves welfare*, because more of a good is supplied in a situation where market valuation exceeds marginal cost.

In the preceding example we simplified the demand side of the market by assuming that every individual in a given group was willing to pay the same price – either \$12 or \$5. More realistically each group can be defined by a downward-sloping demand curve, reflecting the variety of prices that buyers in a given market segment are willing to pay. It is valuable to extend the analysis to include this reality. For example, a supplier may face different demands from her domestic and foreign buyers, and if she can segment these markets she can price discriminate effectively.

Consider Figure 10.12 where two segmented demands are displayed, D_A and D_B, with their associated marginal revenue curves, MR_A and MR_B. We will assume that marginal costs are constant for the moment. It should be clear by this point that the profit maximizing solution for the monopoly supplier is to supply an amount to each market where the MC equals the MR in each market: Since the buyers in one market cannot resell to buyers in the other, the monopolist considers these as two different markets and therefore maximizes profit by applying the standard rule. She will maximize profit in market A by supplying the quantity Q_A and in market B by supplying Q_B. The prices at which these quantities can be sold are P_A and P_B. These prices, unsurprisingly, are different – the objective of segmenting markets is to increase profit by treating the markets as distinct.

An example of this type of price discrimination is where pharmaceutical companies sell drugs to less developed economies at a lower price than to developed economies. The low price is sufficient to cover marginal cost and is therefore profitable - provided the high price market covers the fixed costs.

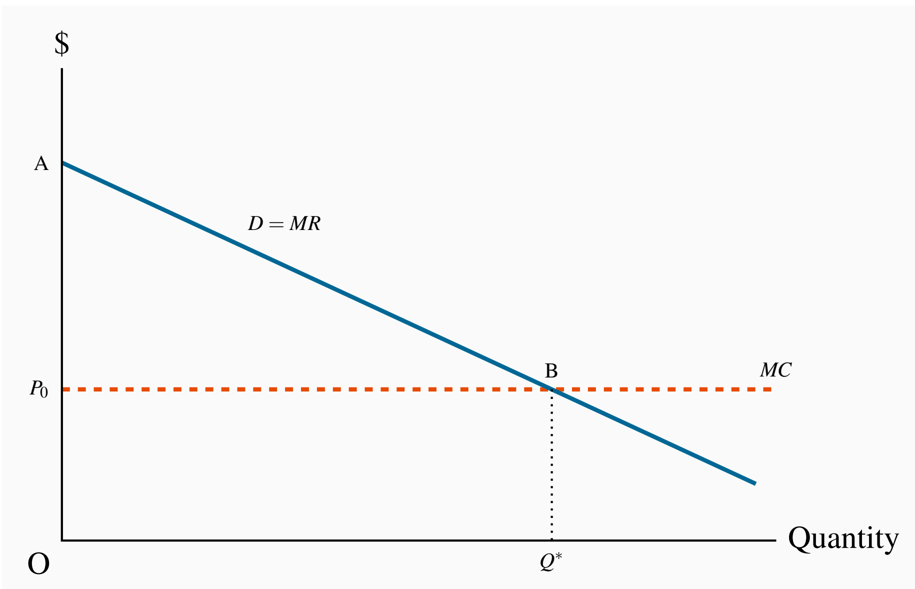
Figure 10.12: Pricing in segregated markets



With two separate markets defined by D_A and D_B, and their associated MR curves MR_A and MR_B, a profit maximizing strategy is to produce where $MC = MR_A = MR_B$, and *discriminate* between the two markets by charging prices P_A and P_B.

The preceding examples involved two separable groups of customers and are very real. This kind of group segregation is sometimes called *third degree price discrimination*. But it may be possible to segregate customers into several groups rather than just two. In the limit, if we could charge a different price to every consumer in a market, or for every unit sold, the revenue accruing to the monopolist would be the area under the demand curve up to the output sold. Though primarily of theoretical interest, this is illustrated in Figure 10.13. It is termed *perfect price discrimination*, and sometimes *first degree price discrimination*. Such discrimination is not so unrealistic: A tax accountant may charge different customers a different price for providing the same service; home renovators may try to charge as much as any client appears willing to pay.

Figure 10.13: Perfect price discrimination



A monopolist who can sell each unit at a different price maximizes profit by producing Q^*. With each consumer paying a different price the demand curve becomes the *MR* curve. The result is that the monopoly DWL is eliminated because the efficient output is produced, and the monopolist appropriates all the consumer surplus. Total revenue for the perfect price discriminator is OABQ^*.

Second degree price discrimination is based on a different concept of buyer identifiability. In the cases we have developed above, the seller is able to distinguish the buyers by *observing* a vital characteristic that signals their type. It is also possible that, while individuals might have defining traits which influence their demands, such traits might not be detectable by the supplier. Nonetheless, it is frequently possible for the supplier to offer different pricing options (corresponding to different uses of a product) that buyers would choose from, with the result that her profit would be greater than under a uniform price with no variation in the use of the service. Different cell phone 'plans', or different internet plans that users can choose from are examples of this second-degree discrimination.

10.6 Cartels: Acting like a monopolist

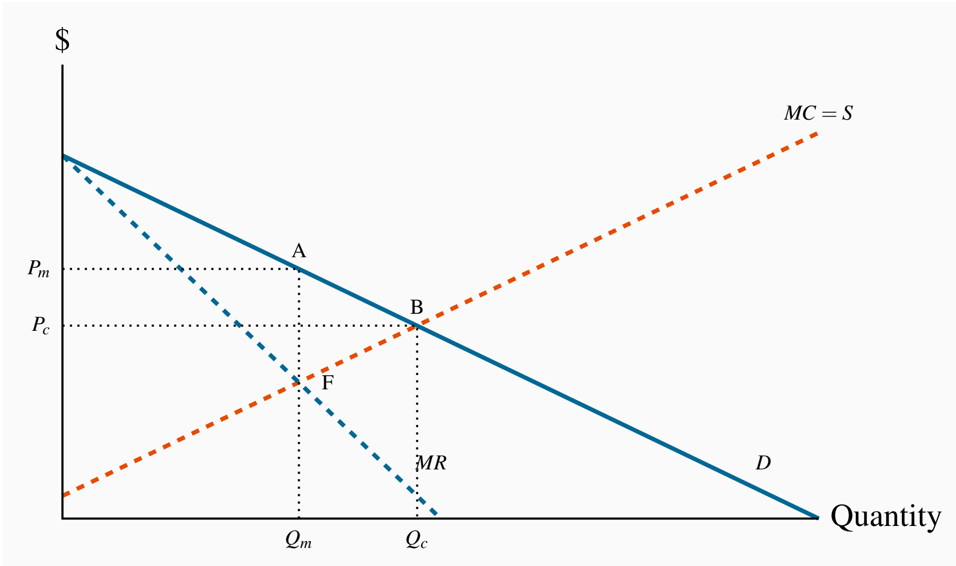
A **cartel** is a group of suppliers that colludes to operate like a monopolist. The cartel formed by the members of the Organization of Oil Exporting Countries (OPEC) is an example of a cartel that was successful in achieving its objectives for a long period. This cartel first flexed its muscles in 1973, by increasing the world price of oil from $3 per barrel to $10 per barrel. The result was to transfer billions of dollars from the energy-importing nations in Europe and North America to OPEC members – the demand for oil is relatively inelastic, hence an increase in price increases total expenditures.

> A **cartel** is a group of suppliers that colludes to operate like a monopolist.

A second renowned cartel is managed by De Beers, which controls a large part of the world's diamond supply. In Canada, agricultural marketing boards are a means of restricting supply legally. Such cartels may have thousands of members. By limiting entry, through requiring a production 'quota', the incumbents can charge a higher price than if entry to the industry were free.

To illustrate the dynamics of cartels consider Figure 10.14. Several producers, with given production capacities, come together and agree to restrict output with a view to increasing price and therefore profit. This may be done with the agreement of the government, or it may be done secretively, and possibly against the law. Each firm has a MC curve, and the industry supply is defined as the sum of these marginal cost curves, as illustrated in Figure 9.3. The resulting cartel is effectively one in which there is a single supplier with many different plants – a multi-plant monopolist. To maximize profits this organization will choose an output level Q_m where the MR equals the MC. In contrast, if these firms act competitively the output chosen will be Q_c. The competitive output yields no supernormal profit, whereas the monopoly/cartel output does.

Figure 10.14: Cartelizing a competitive industry



A cartel is formed when individual suppliers come together and act like a monopolist in order to increase profit. If MC is the joint supply curve of the cartel, profits are maximized at the output Q_m, where $MC = MR$. In contrast, if these firms operate competitively output increases to Q_c.

The cartel results in a deadweight loss equal to the area ABF, just as in the standard monopoly model.

Cartel instability

Some cartels are unstable in the long run. In the first instance, the degree of instability depends on the authority that the governing body of the cartel can exercise over its members, and upon the degree of information it has on the operations of its members. If a cartel is simply an arrangement among producers to limit output, each individual member of the cartel has an incentive to increase its output, because the monopoly price that the cartel attempts to sustain exceeds the cost of producing a marginal unit of output. In Figure 10.14 each firm has a MC of output equal to \$F when the group collectively produces the output Q_m. Yet any firm that brings output to market, *beyond its agreed production limit*, at the price P_m will make a profit of AF on that additional output – *provided the other members of the cartel agree to restrict their output*. Since each firm faces the same incentive to increase output, it is difficult to restrain all members from doing so.

Individual members are more likely to abide by the cartel rules if the organization can sanction them for breaking the supply-restriction agreement. Alternatively, if the actions of individual members are not observable by the organization, then the incentive to break ranks may be too strong for the cartel to sustain its monopoly power.

We will see in Chapter 14 that Canada's Competition Act forbids the formation of cartels, as it forbids many other anti-competitive practices. At the same time, our governments frequently are the driving force in the formation of domestic cartels.

In the second instance, cartels may be undermined eventually by the emergence of new products and new technologies. OPEC has lost much of its power in the modern era because of technological developments in oil recovery. Canada's 'tar sands' yield oil, as a result of technological developments that enabled producers to separate the oil from the earth it is mixed with. Fracking technologies are another means of extracting oil that is discovered in small pockets and encased in rock. The supply coming from these new technologies has limited the ability of the old OPEC cartel to increase prices through supply restriction.

Application Box 10.1: The taxi cartel

The new sharing economy has brought competition to some traditional cartels. City taxis are an example of such a formation: Traditionally, entry has been restricted to drivers who hold a permit (medallion), and fares are higher as a consequence of the resulting reduced supply. A secondary market then develops for these medallions, in which the city may offer new medallions through auction, or existing owners may exit and sell their medallions. Restricted entry has characterized most of Canada's major cities. Depending on the strictness of the entry process, medallions are worth correspondingly more. By 2012, medallions were selling in New York and Boston for a price in the neighborhood of one million dollars.

But ride-sharing start-up companies changed all of that. As Western examples, Uber and Lyft developed smart-phone apps that link demanders for rides with drivers, who may, or may not be, part of the traditional taxi companies. Such start-ups have succeeded in taking a significant part of the taxi business away from the traditional operators. As a result, the price of taxi medallions on the open market has plunged. From trading in the range of $1m. in New York in 2012, medallions are being offered in 2019 at about one fifth of that price. In Toronto, some medallions were traded in the range of $300,000 in 2012, but are on offer in 2019 for prices in the range of $30,000.

Not surprisingly, the traditional taxi companies charge that ride-hailing operators are violating the accepted rules governing the taxi business, and have launched legal suits against them and against local governments, and lobbied governments to keep them out of their cities.

In the new 'sharing economy', of which ride hailing companies are an example, participants operate with less traditional capital, and the communications revolution has been critical to their success. Home owners can use an online site to rent a spare bedroom in their house to visitors to their city (*Airbnb*), and thus compete with hotels. The main capital in this business is in the form of the information technology that links potential buyers to potential sellers.

Information on medallion prices in Canada can be found by, for example, searching at `http://www.kijiji.ca`

10.7 Invention, innovation and rent seeking

Invention and innovation are critical aspects of the modern economy. In some sectors of the economy, firms that cannot invent or innovate are liable to die. Invention is a genuine discovery, whereas innovation is the introduction of a new product or process.

> **Invention** is the discovery of a new product or process through research.
>
> **Product innovation** refers to new or better goods or services.
>
> **Process innovation** refers to new or better production or supply.

To this point we have said little that is good about monopolies. However, the economist Joseph Schumpeter argued that, while monopoly leads to resource misallocation in the economy, this cost might be offset by the greater tendency for monopoly firms to invent and innovate. This is because such firms have more profit and therefore more resources with which to fund R&D and may therefore be more innovative than competitive firms. If this were true then, taking a long-run dynamic view of the marketplace, monopolies could have lower costs and more advanced products than competitive firms and thus benefit the consumer.

While this argument has some logical appeal, it falls short on several counts. First, even if large firms carry out more research than competitive firms, there is no guarantee that the ensuing benefits carry over to the consumer. Second, the results of such research may be used to prevent entry into the industry in question. Firms may register their inventions and gain use protection before a competitor can come up with the same or a similar invention. *Apple* and *Samsung* each own tens of thousands of patents. Third, the empirical evidence on the location of most R&D is inconclusive: A sector with several large firms, rather than one with a single or very many firms, may be best. For example, if *Apple* did not have *Samsung* as a competitor, or vice versa, would the pace of innovation be as strong?

Fourth, much research has a 'public good' aspect to it. Research carried out at universities and government-funded laboratories is sometimes referred to as basic research: It explores the principles underlying chemistry, social relations, engineering forces, microbiology, etc., and has multiple applications in the commercial world. If disseminated, this research is like a public good – its fruits can be used in many different applications, and its use in one area does not preclude its use in others. Consequently, rather than protecting monopolies on the promise of more R&D, a superior government policy might be to invest directly in research and make the fruits of the research publicly available.

Modern economies have **patent laws**, which grant inventors a legal monopoly on use for a fixed period of time – perhaps fifteen years. By preventing imitation, patent laws raise the incentive to conduct R&D but do not establish a monopoly in the long run. Over the life of a patent the inventor charges a higher price than would exist if his invention were not protected; this both yields greater profits and provides the research incentive. When the patent expires, competition from other producers leads to higher output and lower prices for the product. Generic drugs are a good example of this phenomenon.

> **Patent laws** grant inventors a legal monopoly on use for a fixed period of time.

The power of globalization once again is very relevant in patents. Not all countries have patent laws that are as strong as those in North America and Europe. The BRIC economies (Brazil, Russia, India and China) form an emerging power block. But their legal systems and enforcement systems are less well-developed than in Europe or North America. The absence of a strong and transparent legal structure inhibits research and development, because their fruits may be appropriated by competitors.

Rent seeking

Citizens are frequently appalled when they read of lobbying activities in their nation's capital. Every capital city in the world has an army of lobbyists, seeking to influence legislators and regulators. Such individuals are in the business of **rent seeking**, whose goal is to direct profit to particular groups, and protect that profit from the forces of competition. In Virginia and Kentucky we find that state taxes on cigarettes are the lowest in the US – because the tobacco leaf is grown in these states, and the tobacco industry makes major contributions to the campaigns of some political representatives.

Rent-seeking carries a resource cost: Imagine that we could outlaw the lobbying business and put these lobbyists to work producing goods and services in the economy instead. Their purpose is to maintain as much quasi-monopoly power in the hands of their clients as possible, and to ensure that the fruits of this effort go to those same clients. If this practice could be curtailed then the time and resources involved could be redirected to other productive ends.

> **Rent seeking** is an activity that uses productive resources to redistribute rather than create output and value.

Industries in which rent seeking is most prevalent tend to be those in which the potential for economic profits is greatest – monopolies or near-monopolies. These, therefore, are the industries that allocate resources to the preservation of their protected status. We do not observe laundromat owners or shoe-repair businesses lobbying in Ottawa.

Conclusion

We have now examined two extreme types of market structure – perfect competition and monopoly. While many sectors of the economy operate in a way that is close to the competitive paradigm, very few are pure monopolies in that they have no close substitute products. Even firms like *Microsoft*, or *De Beers*, that supply a huge percentage of the world market for their product would deny that they are monopolies and would argue that they are subject to strong competitive pressures from smaller or ‘fringe’ producers. As a result we must look upon the monopoly paradigm as a useful way of analyzing markets, rather than being an exact description of the world. Accordingly, our next task is to examine how sectors with a few, several or multiple suppliers act when pursuing the objective of profit maximization. Many different market structures define the real economy, and we will concentrate on a limited number of the more important structures in the next chapter.

Key Terms

Monopolist: is the sole supplier of an industry's output, and therefore the industry and the firm are one and the same.

Natural monopoly: one where the ATC of producing any output declines with the scale of operation.

Marginal revenue is the change in total revenue due to selling one more unit of the good.

Average revenue is the price per unit sold.

Allocative inefficiency arises when resources are not appropriately allocated and result in deadweight losses.

Price discrimination involves charging different prices to different consumers in order to increase profit.

A **cartel** is a group of suppliers that colludes to operate like a monopolist.

Rent seeking is an activity that uses productive resources to redistribute rather than create output and value.

Invention is the discovery of a new product or process through research.

Product innovation refers to new or better products or services.

Process innovation refers to new or better production or supply.

Patent laws grant inventors a legal monopoly on use for a fixed period of time.

Exercises for Chapter 10

Exercise 10.1 Consider a monopolist with demand curve defined by $P = 100 - 2Q$. The MR curve is $MR = 100 - 4Q$ and the marginal cost is $MC = 10 + Q$. The demand intercepts are $\{\$100, 50\}$, the MR intercepts are $\{\$100, 25\}$.

(a) Develop a diagram that illustrates this market, using either graph paper or an Excel spreadsheet, for values of output $Q = 1 \ldots 25$.

(b) Identify visually the profit-maximizing price and output combination.

(c) *Optional*: Compute the profit maximizing price and output combination.

Exercise 10.2 Consider a monopolist who wants to maximize revenue rather than profit. She has the demand curve $P = 72 - Q$, with marginal revenue $MR = 72 - 2Q$, and $MC = 12$. The demand intercepts are $\{\$72, 72\}$, the MR intercepts are $\{\$72, 36\}$.

(a) Graph the three functions, using either graph paper or an Excel spreadsheet.

(b) Calculate the price she should charge in order to maximize revenue. [*Hint*: Where the $MR = 0$.]

(c) Compute the total revenue she will obtain using this strategy.

Exercise 10.3 Suppose that the monopoly in Exercise 10.2 has a large number of plants. Consider what could happen if each of these plants became a separate firm, and acted competitively. In this perfectly competitive world you can assume that the MC curve of the monopolist becomes the industry supply curve.

(a) Illustrate graphically the output that would be produced in the industry?

(b) What price would be charged in the marketplace?

(c) *Optional*: Compute the gain to the economy in dollar terms as a result of the DWL being eliminated [*Hint*: It resembles the area ABF in Figure 10.14].

Exercise 10.4 In the text example in Table 10.1, compute the profit that the monopolist would make if he were able to price discriminate, by selling each unit at the demand price in the market.

Exercise 10.5 A monopolist is able to discriminate perfectly among his consumers – by charging a different price to each one. The market demand curve facing him is given by $P = 72 - Q$. His marginal cost is given by $MC = 24$ and marginal revenue is $MR = 72 - 2Q$.

(a) In a diagram, illustrate the profit-maximizing equilibrium, where discrimination is not practiced. The demand intercepts are $\{\$72, 72\}$, the MR intercepts are $\{\$72, 36\}$.

(b) Illustrate the equilibrium output if he discriminates perfectly.

(c) *Optional*: If he has no fixed cost beyond the marginal production cost of \$24 per unit, calculate his profit in each pricing scenario.

Exercise 10.6 A monopolist faces two distinct markets A and B for her product, and she is able to insure that resale is not possible. The demand curves in these markets are given by $P_A = 20 - (1/4)Q_A$ and $P_B = 14 - (1/4)Q_B$. The marginal cost is constant: $MC = 4$. There are no fixed costs.

(a) Graph these two markets and illustrate the profit maximizing price and quantity in each market. [You will need to insert the *MR* curves to determine the optimal output.] The demand intercepts in A are $\{\$20, 80\}$, and in B are $\{\$14, 56\}$.

(b) In which market will the monopolist charge a higher price?

Exercise 10.7 A concert organizer is preparing for the arrival of the Grateful Living band in his small town. He knows he has two types of concert goers: One group of 40 people, each willing to spend \$60 on the concert, and another group of 70 people, each willing to spend \$40. His total costs are purely fixed at \$3,500.

(a) Draw the market demand curve faced by this monopolist.

(b) Draw the *MR* and *MC* curves.

(c) With two-price discrimination what will be the monopolist's profit?

(d) If he must charge a single price for all tickets can he make a profit?

Exercise 10.8 *Optional*: A monopolist faces a demand curve $P = 64 - 2Q$ and $MR = 64 - 4Q$. His marginal cost is $MC = 16$.

(a) Graph the three functions and compute the profit maximizing output and price.

(b) Compute the efficient level of output (where *MC*=demand), and compute the DWL associated with producing the profit maximizing output rather than the efficient output.

Chapter 11

Imperfect competition

In this chapter we will explore:

11.1 The principle ideas

The preceding chapters have explored extreme forms of supply: The monopolist is the sole supplier and possesses as much market power as possible. In contrast, the perfect competitor is small and has no market power whatsoever. He simply accepts the price for his product that is determined in the market by the forces of supply and demand. These are very useful paradigms to explore, but the real world for the most part lies between these extremes. We observe that there are a handful of dominant brewers in Canada who supply more than three quarters of the market, and they are accompanied by numerous micro brewers that form the fringe of the brewing business. We have a small number of air carriers and one of them controls half of the national market. The communications market has just three major suppliers; the Canadian Football League has nine teams and there are just a handful of major hardware/builders' suppliers stores nationally. At the other end of the spectrum we have countless restaurants and fitness centres, but they do not supply exactly the same product to the marketplaces for 'food' or 'health', and so these markets are not perfectly competitive, despite the enormous number of participants.

In this chapter we will explore three broad topics: First is the relationship between firm behaviour and firm size relative to the whole sector. This comes broadly under the heading of *imperfect* *competition* and covers a variety of market forms. Second, we will explore the principle modern ideas in *strategic behavior*. In a sense all decisions in microeconomics have an element of strategy to them - economic agents aim to attain certain goals and they adopt specific maximizing strategies to attain them. But in this chapter we explore a more specific concept of strategic behavior - one that focuses upon direct interactions between a small number of players in the market place. Third, we explore the principle characteristics of what are termed *matching' markets*. These are markets where transactions take place without money and involve matching heterogeneous suppliers with heterogeneous buyers.

11.2 Imperfect competitors

Imperfect competitors can be defined by the number of firms in their sector, or the share of total sales going to a small number of suppliers. They can also be defined in terms of the characteristics of the demand curves they all face. A perfect competitor faces a perfectly elastic demand at the existing market price, and this is the only market structure to have this characteristic. In all other market structures suppliers effectively face a downward-sloping demand. This means that they have some influence on the price of the good, and also that if they change the price they charge, they can expect demand to reflect this in a predictable manner. So, in theory, we can classify all market structures apart from perfect competition as being **imperfectly competitive**. In practice we use the term to denote firms that fall between the extremes of perfect competition and monopoly.

> **Imperfectly competitive firms** face a downward-sloping demand curve, and their output price reflects the quantity sold.

The demand curve for the firm and industry coincide for the monopolist, but not for other imperfectly competitive firms. It is convenient to categorize the producing sectors of the economy as either having a relatively small number of participants, or having a large number. The former market structures are called **oligopolistic**, and the latter are called **monopolistically competitive**. The word *oligopoly* comes from the Greek word *oligos* meaning few, and *polein* meaning to sell.

> **Oligopoly** defines a market with a small number of suppliers.
>
> **Monopolistic competition** defines a market with many sellers of products that have similar characteristics. Monopolistically competitive firms can exert only a small influence on the whole market.

The home appliance industry is an oligopoly. The prices of *KitchenAid* appliances depend not only on their own output and sales, but also on the prices of *Whirlpool*, *Maytag* and *Bosch*. If a firm has just two main producers it is called a **duopoly**. *Canadian National* and *Canadian Pacific* are the only two major rail freight carriers in Canada; they thus form a duopoly. In contrast, the local Italian restaurant is a monopolistic competitor. Its output is a package of distinctive menu choices, personal service, and convenience for local customers. It can charge a different price than the out-of-neighbourhood restaurant, but if its prices are too high local diners may travel elsewhere for their food experience, or switch to a different cuisine locally. Many markets are defined by producers who supply similar but not identical products. Canada's universities all provide degrees, but they differ one from another in their programs, their balance of in-class and on-line courses, their student activities, whether they are science based or liberal arts based, whether they have cooperative programs or not, and so forth. While universities are not in the business of making profit, they certainly wish to attract students, and one way of doing this is to differentiate themselves from other institutions. The profit-oriented world of commerce likewise seeks to increase its market share by distinguishing its product line.

Duopoly defines a market or sector with just two firms.

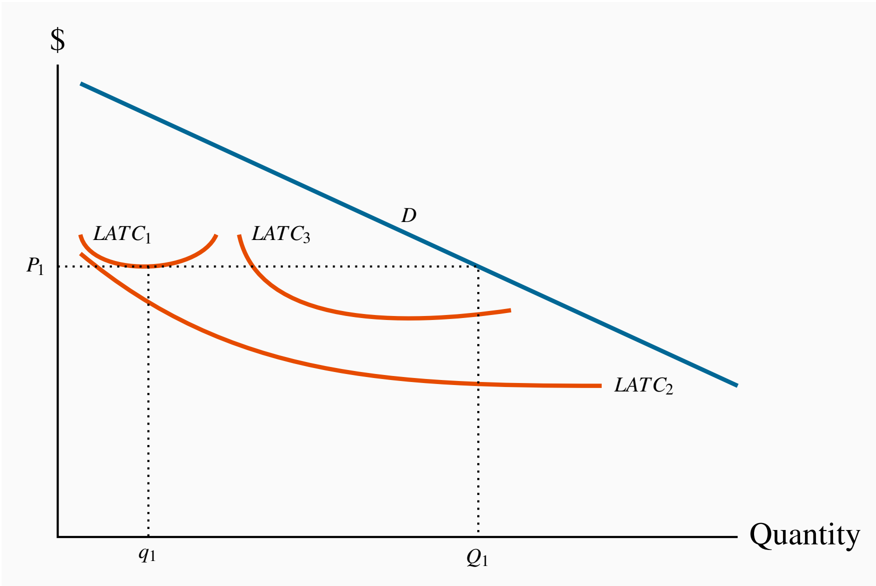
These distinctions are not completely airtight. For example, if a sole domestic producer is subject to international competition it cannot act in the way we described in the previous chapter – it has potential, or actual, competition. *Bombardier* may be Canada's sole rail car manufacturer, but it is not a monopolist, even in Canada. It could best be described as being part of an international oligopoly in rail-car manufacture. Likewise, it is frequently difficult to delineate the boundary of a given market. For example, is *Canada Post* a monopoly in mail delivery, or an oligopolist in hard-copy communication? We can never fully remove these ambiguities.

The role of cost structures

A critical determinant of market structure is the way in which demand and cost interact to determine the likely number of market participants in a given sector or market. Structure also evolves over the long run: Time is required for entry and exit.

Figure 11.1 shows the demand curve D for the output of an industry in the long run. Suppose, initially, that all firms and potential entrants face the long-run average cost curve $LATC_1$. At the price P_1, free entry and exit means that each firm produces q_1. With the demand curve D, industry output is Q_1. The number of firms in the industry is N_1 $(= Q_1/q_1)$. If q_1, the minimum average cost output on $LATC_1$, is small relative to D, then N_1 is large. This outcome might be perfect competition (N virtually infinite), or monopolistic competition (N large) with slightly differentiated products produced by each firm.

Figure 11.1: Demand, costs and market structure



With a cost structure defined by $LATC_1$ this market has space for many firms – perfect or monopolistic competition, each producing approximately q_1. If costs correspond to $LATC_2$, where scale economies are substantial, there may be space for just one producer. The intermediate case, $LATC_3$, can give rise to oligopoly, with each firm producing more than q_1 but less than a monopolist. These curves encounter their MES at very different output levels.

Instead, suppose that the production structure in the industry is such that the long-run average cost curve is $LATC_2$. Here, scale economies are vast, relative to the market size. At the lowest point on this cost curve, output is large relative to the demand curve D. If this one firm were to act like a monopolist it would produce an output where $MR = MC$ in the long run and set a price such that the chosen output is sold. Given the scale economies, there may be no scope for another firm to enter this market, because such a firm would have to produce a very high output to compete with the existing producer. This situation is what we previously called a “natural” monopolist.

Finally, the cost structure might involve curves of the type $LATC_3$, which would give rise to the possibility of several producers, rather than one or very many. This results in oligopoly.

It is clear that one *crucial determinant of market structure is minimum efficient scale relative to the size of the total market* as shown by the demand curve. The larger the minimum efficient scale relative to market size, the smaller is the number of producers in the industry.

11.3 Imperfect competitors: measures of structure and market power

Sectors of the economy do not fit neatly into the limited number of categories described above. The best we can say in most cases is that they resemble more closely one type of market than another. Consider the example of Canada's brewing sector: It has two large brewers in *Molson-Coors* and *Labatt*, a couple of intermediate sized firms such as *Sleeman*, and an uncountable number of small boutique brew pubs. While such a large number of brewers satisfy one requirement for perfect competition, it would not be true to say that the biggest brewers wield no market power; and this is the most critical element in defining market structure.

By the same token, we could not define this market as a duopoly: Even though there are just two major participants, there are countless others who, together, are important.

One way of defining what a particular structure most closely resembles is to examine the percentage of sales in the market that is attributable to a small number of firms. For example: What share is attributable to the largest three or four firms? The larger the share, the more concentrated the market power. Such a statistic is called a concentration ratio. The **N-firm concentration ratio** is the sales share of the largest N firms in that sector of the economy.

> The **N-firm concentration ratio** is the sales share of the largest N firms in that sector of the economy.

Table 11.1: Concentration in Canadian food processing 2011

Sector	% of shipments
Sugar	98
Breakfast cereal	96
Canning	60
Meat processing	23

Source: "Four Firm Concentration Ratios (CR4s) for selected food processing sectors," adapted from Statistics Canada publication Measuring industry concentration in Canada's food processing sectors, Agriculture and Rural Working Paper series no. 70, Catalogue 21-601, `http://www.statcan.gc.ca/pub/21-601-m/21-601-m2004070-eng.pdf`.

Table 11.1 contains information on the 4-firm concentration ratio for several sectors of the Canadian economy. It indicates that, at one extreme, sectors such as breakfast cereals and sugars have a high degree of concentration, whereas meat processing has much less. A high degree of concentration suggests market power, and possibly economies of scale.

11.4 Imperfect competition: monopolistic competition

Monopolistic competition presumes a large number of quite small producers or suppliers, each of whom may have a slightly **differentiated product**. The competition element of this name signifies that there are many participants, while the monopoly component signifies that each supplier faces a downward-sloping demand. In concrete terms, your local coffee shop that serves "fair trade" coffee has a product that differs slightly from that of neighbouring shops that sell the traditional product. They coexist in the same sector, and probably charge different prices: The fair trade supplier likely charges a higher price, but knows nonetheless that too large a difference between her price and the prices of her competitors will see some of her clientele migrate to those lower-priced establishments. That is to say, she faces a downward-sloping demand curve.

The competition part of the name also indicates that there is *free entry and exit*. There are no barriers to entry. As a consequence, we know at the outset that only normal profits will exist in a long-run equilibrium. Economic profits will be competed away by entry, just as losses will erode due to exit.

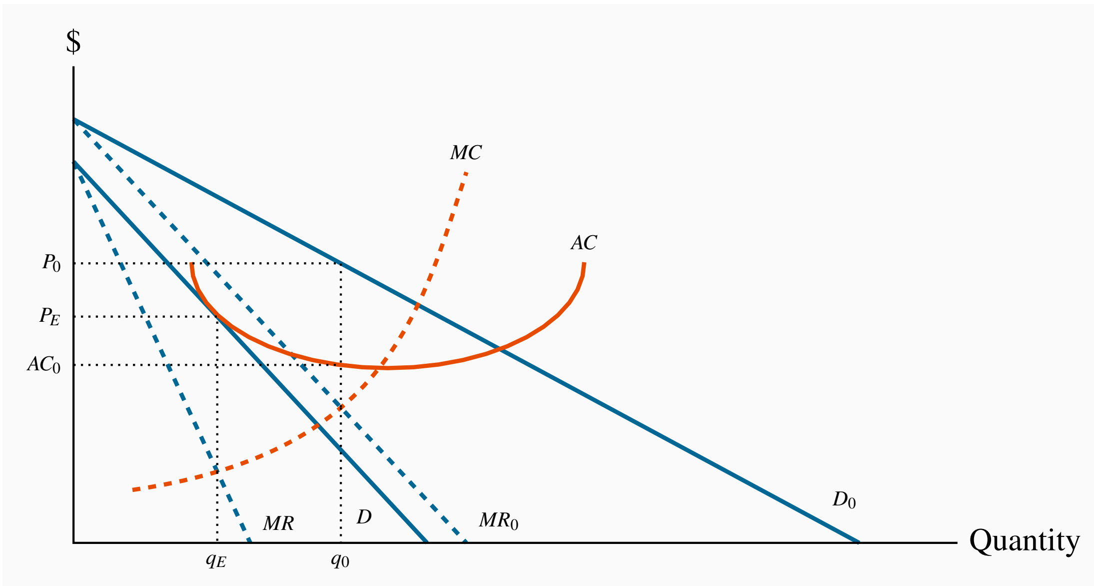
As a general rule then, each firm can influence its market share to some extent by changing its price. Its demand curve is not horizontal because different firms' products are only limited substitutes. A lower price level may draw some new customers away from competitors, but convenience or taste will prevent most patrons from deserting their local businesses. In concrete terms: A pasta special at the local Italian restaurant that reduces the price below the corresponding price at the competing local Thai restaurant will indeed draw clients away from the latter, but the foods are sufficiently different that only some customers will leave the Thai restaurant. The differentiated menus mean that many customers will continue to pay the higher price.

> A **differentiated product** is one that differs slightly from other products in the same market.

Given that there are very many firms, the theory also envisages limits to scale economies. Firms are small and, with many competitors, individual firms do not compete strategically with *particular* rivals. Because the various products offered are slightly differentiated, we avoid graphics with a *market* demand, because this would imply that a uniform product is being considered. At the same time the market is a well-defined concept—it might be composed of all those restaurants within a reasonable distance, for example, even though each one is slightly different from the others. The market share of each firm depends on the price that it charges *and* on the number of competing firms. For a given number of suppliers, a shift in industry demand also shifts the demand facing each firm. Likewise, the presence of more firms in the industry reduces the demand facing each one.

Equilibrium is illustrated in Figure 11.2. Here D_0 is the initial demand facing a representative firm, and MR_0 is the corresponding marginal revenue curve. Profit is maximized where $MC = MR$, and the price P_0 is obtained from the demand curve corresponding to the output q_0. Total profit is the product of output times the difference between price and average cost, which equals $q_0 \times (P_0 - AC_0)$.

Figure 11.2: Equilibrium for a monopolistic competitor



Profits exist at the initial equilibrium (q_0, P_0). Hence, new firms enter and reduce the share of the total market faced by each firm, thereby shifting back their demand curve. A final equilibrium is reached where economic profits are eliminated: At $AC = P_E$ and $MR = MC$.

With free entry, such profits attract new firms. The increased number of firms reduces the share of the market that any one firm can claim. That is, the firm's demand curve shifts inwards when entry occurs. As long as (economic) profits exist, this process continues. For entry to cease, average cost must equal price. A final equilibrium is illustrated by the combination (P_E, q_E), where the demand has shifted inward to D.

At this long-run equilibrium, two conditions must hold: First, the optimal pricing rule must be satisfied—that is $MC = MR$; second it must be the case that only normal profits are made at the final equilibrium. Economic profits are competed away as a result of free entry. Graphically this implies that ATC must equal price at the output where $MC = MR$. In turn this implies that the ATC is tangent to the demand curve where $P = ATC$. While this could be proven mathematically, it is easy to intuit why this tangency must exist: If ATC merely intersected the demand curve at the output where $MC = MR$, we could find some other output where the demand price would be above ATC, suggesting that profits could be made at such an output. Clearly that could not represent an equilibrium.

The **monopolistically competitive equilibrium** in the long run requires the firm's demand curve to be tangent to the ATC curve at the output where $MR = MC$.

11.5 Imperfect competition: economies of scope and platforms

The communications revolution has impacted market structure in modern economies profoundly: it has facilitated economies of scope, meaning that firms may yield more collective profit if merged than if operating independently.

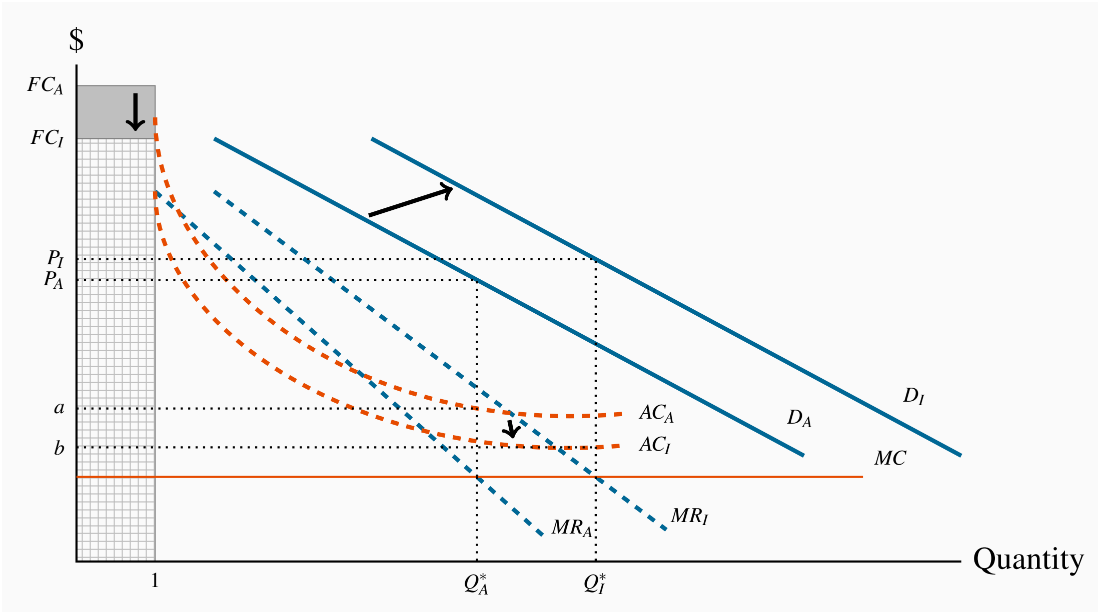
Economies of Scope

Imagine an aspiring entrepreneur who envisages a revolution of the traditional taxi sector of the economy. He decides to develop a smartphone application that will match independent income-seeking vehicle owners (drivers) with individuals seeking transport (passengers) from point A to point B. We know how this adventure evolves. In one case it takes the form of the corporation Uber, in another the corporation Lyft, and others worldwide.

These corporations have grown in leaps and bounds and have taken business from the conventional taxi corporations. As of 2019 they cannot turn a profit, yet the stock market continues to bet upon future success: investors believe that when these corporations evolve into fully integrated multi-product suppliers, both costs will decline and demand will increase for each component of the business. In the case of transportation companies, they aim to become a 'one-stop-shop' for mobility services. Uber is not only a ride-hailing service, it also transports meals through its Uber-eats platform, and is developing the electric scooter and electric bike markets in addition. In some local markets it is linked to public transport services. All of this is being achieved through a single smartphone application. The objective is to simplify movement for persons, by providing multiple options on a variety of transport modes, accessed through a single portal.

This phenomenon is described in Figure 11.3. The subscripts A and I represent market conditions when the service supplier is operating Alone or in an Integrated corporation. The initial equilibrium is defined by the A demand and cost conditions. The profit maximizing output occurs when $MC_A = MR_A$, leading to a price P_A and a quantity Q_A^*. Each unit of the good yields a profit margin of $(P_A - a)$.

Figure 11.3: Summa's ride hailing service



Demand for a particular product increases when the autonomous supplier (A) merges with another firm to become an integrated firm (I), because customers switch to firms that offer several different services from the same platform: the demand curve shifts outward, from D_A to D_I. With integration, the fixed costs fall and average costs fall, even with marginal costs constant. Output and profit increase, and concentration in the marketplace rises.

This firm now merges with another transportation corporation - perhaps a food delivery service, perhaps an electric bike service. Since each firm has a similar type of fixed cost, these costs can be reduced by the merger. In technical terms, the merged firms, or merged operations, share a common hardware-cum-software **platform**. Each firm will therefore incur lower average costs, even if marginal costs remain unchanged: the AC curve declines to AC_I. In addition to the decline in average costs, each firm sees an increase in its customer base, because transportation service buyers find it preferable to choose their mode of transport through a single portal rather than through several different modes of access. This is represented by an outward shift in the demand curve for vehicle rides to D_I.

The new profit maximizing equilibrium occurs at Q_I^*. Total profit necessarily increases both because average costs have fallen and the number of buyers willing to buy at any price has risen. The analytics in this figure also describe the benefits accruing to the other firm or firms in the merger.

> A **platform** describes a technology that is common to more than one product in a multi-product organization.

We conclude from this analysis that, if scope economies are substantial, it may be difficult for stand-alone firms specializing in just one component of the transportation services sector to remain profitable. It may also be impossible to define a conventional equilibrium in this kind of marketplace. This is because some conglomerate firms may have different component producers in their suite of firms. For example, Lyft may not have a food delivery service, but it may have a limousine or bus service. What is critical for an equilibrium is that firms of a particular type, whether they are part of a conglomerate or not, be able to compete with corresponding firms. This means that their cost structure must be similar.

As a further example: Amazon initially was primarily an on-line book seller. But it expanded to include the sale of other products. And once it became a 'market for everything' the demand side of the market exploded in parallel with the product line, because it becomes easy to shop for 'anything' or even different objects on a single site. Only Walmart, in North America, comes close to being able to compete with Amazon.

11.6 Strategic behaviour: Oligopoly and games

Under perfect competition or monopolistic competition, there are so many firms in the industry that each one can ignore the immediate effect of its own actions on particular rivals. However, in an oligopolistic industry *each firm must consider how its actions affect the decisions of its relatively few competitors*. Each firm must guess how its rivals will react. Before discussing what constitutes an intelligent guess, we investigate whether they are likely to collude or compete. **Collusion** is a means of reducing competition with a view to increasing profit.

> **Collusion** is an explicit or implicit agreement to avoid competition with a view to increasing profit.

A particular form of collusion occurs when firms co-operate to form a cartel, as we saw in the last chapter. Collusion is more difficult if there are many firms in the industry, if the product is not standardized, or if demand and cost conditions are changing rapidly. In the absence of collusion, each firm's demand curve depends upon how competitors react: If Air Canada contemplates offering customers a seat sale on a particular route, how will West Jet react? Will it, too, make the same offer to buyers? If Air Canada thinks about West Jet's likely reaction, will it go ahead with the contemplated promotion? A **conjecture** is a belief that one firm forms about the strategic reaction of another competing firm.

> A **conjecture** is a belief that one firm forms about the strategic reaction of another competing firm.

Good poker players will attempt to anticipate their opponents' moves or reactions. Oligopolists are like poker players, in that they try to anticipate their rivals' moves. To study interdependent decision making, we use game theory. A **game** is a situation in which contestants plan strategically to maximize their payoffs, taking account of rivals' behaviour.

> A **game** is a situation in which contestants plan strategically to maximize their payoffs, taking account of rivals' behaviour.

The *players* in the game try to maximize their own *payoffs*. In an oligopoly, the firms are the players and their payoffs are their profits. Each player must choose a **strategy**, which is a plan describing how a player moves or acts in different situations.

> A **strategy** is a game plan describing how a player acts, or moves, in each possible situation.

Equilibrium outcomes

How do we arrive at an equilibrium in these games? Let us begin by defining a commonly used concept of equilibrium. A **Nash equilibrium** is one in which each player chooses the best strategy, given the strategies chosen by the other players, and there is no incentive to move or change choice.

> A **Nash equilibrium** is one in which each player chooses the best strategy, given the strategies chosen by the other player, and there is no incentive for any player to move.

In such an equilibrium, no player wants to change strategy, since the other players' strategies were already figured into determining each player's own best strategy. This concept and theory are attributable to the Princeton mathematician John Nash, who was popularized by the Hollywood movie version of his life, *A Beautiful Mind*.

In most games, each player's best strategy depends on the strategies chosen by their opponents. Occasionally, a player's best strategy is independent of those chosen by rivals. Such a strategy is called a **dominant strategy**.

> A **dominant strategy** is a player's best strategy, independent of the strategies adopted by rivals.

We now illustrate these concepts with the help of two different games. These games differ in their outcomes and strategies. Table 11.2 contains the domestic happiness game[1]. Will and Kate are attempting to live in harmony, and their happiness depends upon each of them carrying out domestic chores such as shopping, cleaning and cooking. The first element in each pair defines Will's outcome, the second Kate's outcome. If both contribute to domestic life they each receive a happiness or utility level of 5 units. If one contributes and the other does not the happiness levels are 2 for the contributor and 6 for the non-contributor, or 'free-rider'. If neither contributes

[1] This presentation is inspired by a similar problem in Ted Bergstrom and Hal Varian's book *"Workouts"*.

happiness levels are 3 each. When each follows the same strategy the payoffs are on the diagonal, when they follow different strategies the payoffs are on the off-diagonal. Since the elements of the table define the payoffs resulting from various choices, this type of matrix is called a **payoff matrix**.

> **A payoff matrix** defines the rewards to each player resulting from particular choices.

So how is the game likely to unfold? In response to Will's choice of a contribute strategy, Kate's utility maximizing choice involves lazing: She gets 6 units by not contributing as opposed to 5 by contributing. Instead, if Will decides to be lazy what is in Kate's best interest? Clearly it is to be lazy also because that strategy yields 3 units of happiness compared to 2 units if she contributes. In sum, Kate's best strategy is to be lazy, regardless of Will's behaviour. So the strategy of not contributing is a *dominant strategy*, in this particular game.

Will also has a dominant strategy – identical to Kate's. This is not surprising since the payoffs are symmetric in the table. Hence, since each has a dominant strategy of not contributing the Nash equilibrium is in the bottom right cell, where each receives a payoff of 3 units. Interestingly, this equilibrium is not the one that yields maximum combined happiness.

Table 11.2: A game with dominant strategies

		Kate's choice	
		Contribute	Laze
Will's choice	Contribute	5,5	2,6
	Laze	6,2	3,3

The first element in each cell denotes the payoff or utility to Will; the second element the utility to Kate.

The reason that the equilibrium yields less utility for each player in this game is that the game is competitive: Each player tends to their own interest and seeks the best outcome conditional on the choice of the other player. This is evident from the $(5,5)$ combination. From this position Kate would do better to defect to the Laze strategy, because her utility would increase[2].

To summarize: This game has a unique equilibrium and each player has a dominant strategy. But let us change the payoffs just slightly to the values in Table 11.3. The off-diagonal elements have changed. The contributor now gets no utility as a result of his or her contributions: Even though the household is a better place, he or she may be so annoyed with the other person that no utility flows to the contributor.

[2]This game is sometimes called the "prisoners' dilemma" game because it can be constructed to reflect a scenario in which two prisoners, under isolated questioning, each confess to a crime and each ends up worse off than if neither confessed.

Table 11.3: A game without dominant strategies

		Kate's choice	
		Contribute	Laze
Will's choice	Contribute	5,5	0,4
	Laze	4,0	3,3

The first element in each cell denotes the payoff or utility to Will; the second element the utility to Kate.

What are the optimal choices here? Starting again from Will choosing to contribute, what is Kate's best strategy? It is to contribute: She gets 5 units from contributing and 4 from lazing, hence she is better contributing. But what is her best strategy if Will decides to laze? It is to laze, because that yields her 3 units as opposed to 0 by contributing. This set of payoffs therefore *contains no dominant strategy* for either player.

As a result of there being no dominant strategy, there arises the possibility of more than one equilibrium outcome. In fact there are two equilibria in this game now: If the players find themselves both contributing and obtaining a utility level of $(5,5)$ it would not be sensible for either one to defect to a laze option. For example, if Kate decided to laze she would obtain a payoff of 4 utils rather than the 5 she enjoys at the $(5,5)$ equilibrium. By the same reasoning, if they find themselves at the (laze, laze) combination there is no incentive to move to a contribute strategy.

Once again, it is to be emphasized that the twin equilibria emerge in a competitive environment. If this game involved cooperation or collusion the players should be able to reach the $(5,5)$ equilibrium rather than the $(3,3)$ equilibrium. But in the competitive environment we cannot say *ex ante* which equilibrium will be attained.

Repeated games

This game illustrates the tension between collusion and competition. While we have developed the game in the context of the household, it can equally be interpreted in the context of a profit maximizing game between two market competitors. Suppose the numbers define profit levels rather than utility as in Table 11.4. The 'contribute' option can be interpreted as 'cooperate' or 'collude', as we described for a cartel in the previous chapter. They collude by agreeing to restrict output, sell that restricted output at a higher price, and in turn make a greater total profit which they split between themselves. The combined best profit outcome $(5,5)$ arises when each firm restricts its output.

Table 11.4: Collusion possibilities

		Firm K's profit	
		Low output	High output
Firm W's profit	Low output	5,5	2,6
	High output	6,2	3,3

The first element in each cell denotes the profit to Firm W; the second element the profit to Firm K.

But again there arises an incentive to defect: If Firm W agrees to maintain a high price and restrict output, then Firm K has an incentive to renege and increase output, hoping to improve its profit through the willingness of Firm W to restrict output. Since the game is symmetric, each firm has an incentive to renege. Each firm has a dominant strategy – high output, and there is a unique equilibrium $(3,3)$.

Obviously there arises the question of whether these firms can find an operating mechanism that would ensure they each generate a profit of 5 units rather than 3 units, while remaining purely self-interested. This question brings us to the realm of **repeated games**. For example, suppose that firms make strategic choices each quarter of the year. If firm K had 'cheated' on the collusive strategy it had agreed with firm W in the previous quarter, what would happen in the following quarter? Would firms devise a strategy so that cheating would not be in the interest of either one, or would the competitive game just disintegrate into an unpredictable pattern? These are interesting questions and have provoked a great deal of thought among game theorists. But they are beyond our scope at the present time.

> **A repeated game** is one that is repeated in successive time periods and where the knowledge that the game will be repeated influences the choices and outcomes in earlier periods.

We now examine what might happen in one-shot games of the type we have been examining, but in the context of many possible choices. In particular, instead of assuming that each firm can choose a high or low output, how would the outcome of the game be determined if each firm can choose an output that can lie anywhere between a high and low output? In terms of the demand curve for the market, this means that the firms can choose some output and price that is consistent with demand conditions: There may be an infinite number of choices. This framing of a game enables us to explore new concepts in strategic behavior.

11.7 Strategic behaviour: Duopoly and Cournot games

The duopoly model that we frequently use in economics to analyze competition between a small number of competitors is fashioned after the ideas of French economist Augustin Cournot. Consequently it has come to be known as the **Cournot duopoly model**. While the maximizing behaviour that is incorporated in this model can apply to a situation with several firms rather than two, we will develop the model with two firms. This differs slightly from the preceding section, where each firm has simply a choice between a high or low output.

The critical element of the Cournot approach is that the firms each determine their optimal strategy – one that maximizes profit – by reacting optimally to their opponent's strategy, which in this case involves their choice of output.

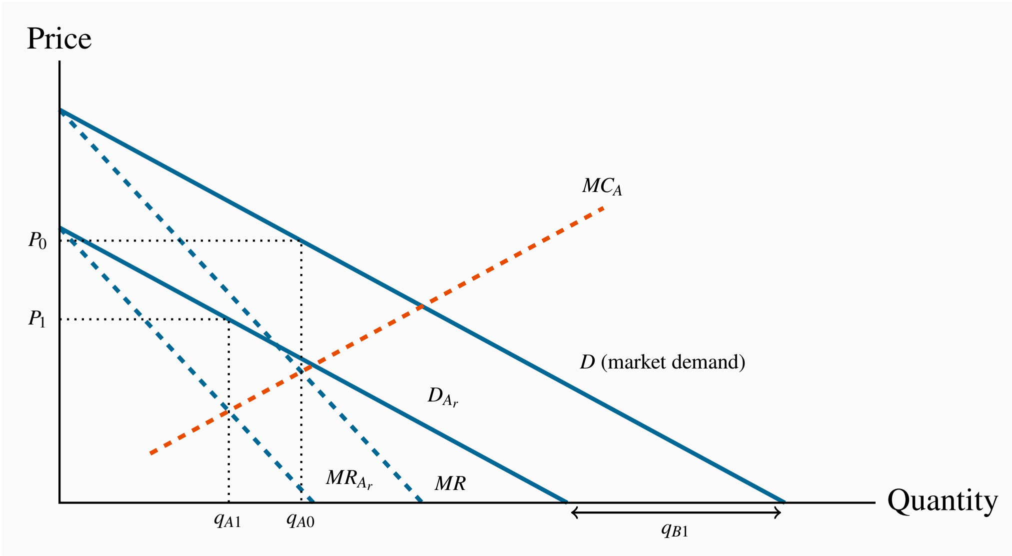
> **Cournot behaviour** involves each firm reacting optimally in their choice of output to their competitors' output decisions.

A central element here is the **reaction function** of each firm, which defines the optimal output choice conditional upon their opponent's choice.

> **Reaction functions** define the optimal choice of output conditional upon a rival's output choice.

We can develop an optimal strategy with the help of Figure 11.4. D is the market demand, and two firms supply this market. If B supplies a zero output, then A would face the whole demand, and would maximize profit where $MC = MR$. Let this output be defined by q_{A0}. We transfer this output combination to Figure 11.5, where the output of each firm is on one of the axes—A on the vertical axis and B on the horizontal. This particular combination of zero output for B and q_{A0} for A is represented on the vertical axis as the point q_{A0}.

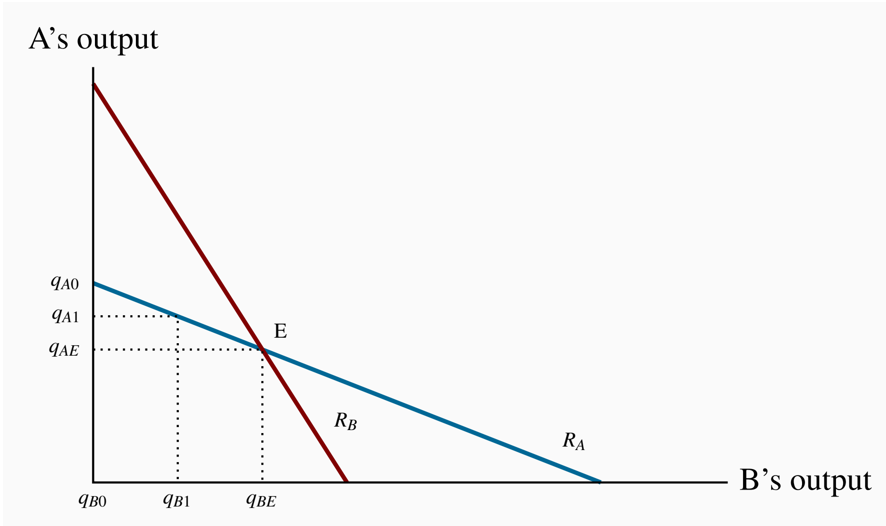
Figure 11.4: Duopoly behaviour



When one firm, B, chooses a specific output, e.g. q_{B1}, then A's residual demand D_{A_r} is the difference between the market demand and q_{B1}. A's profit is maximized at q_{A1} – where $MC = MR_{A_r}$. This is an optimal reaction by A to B's choice. For all possible choices by B, A can form a similar optimal response. The combination of these responses forms A's reaction function.

Instead, suppose that B produces a quantity q_{B1} in Figure 11.4. This reduces the demand curve facing A correspondingly from D to D_{A_r}, which we call A's *residual demand*. When subject to such a choice by B, firm A maximizes profit by producing where $MR_{A_r} = MC$, where MR_{A_r} is the marginal revenue corresponding to the residual demand D_{A_r}. The optimum for A is now q_{A1}, and this pair of outputs is represented by the combination (q_{A1}, q_{B1}) in Figure 11.5.

Figure 11.5: Reaction functions and equilibrium



The reaction function for A (R_A) defines the optimal output response for A to any output choice by B. The reaction function for B is defined similarly. The equilibrium occurs at the intersection of R_A and R_B. Any other combination will induce one firm to change its output, and therefore could not be an equilibrium.

Firm A forms a similar optimal response for every possible output level that B could choose, and these responses define A's reaction function. The reaction function illustrated for A in Figure 11.5 is thus the locus of all optimal response outputs on the part of A. The downward-sloping function makes sense: The more B produces, the smaller is the residual market for A, and therefore the less A will produce.

But A is just one of the players in the game. If B acts in the same optimizing fashion, B too can formulate a series of optimal reactions to A's output choices. The combination of such choices would yield a reaction function for B. This is plotted as R_B in Figure 11.5.

An equilibrium is defined by the intersection of the two reaction functions, in this case by the point E. At this output level *each firm is making an optimal decision, conditional upon the choice of its opponent*. Consequently, neither firm has an incentive to change its output; therefore it can be called the Nash equilibrium.

Any other combination of outputs on either reaction function would lead one of the players to change its output choice, and therefore could not constitute an equilibrium. To see this, suppose that B produces an output greater than q_{BE}; how will A react? A's reaction function indicates that it should choose a quantity to supply less than q_{AE}. If so, how will B respond in turn to that optimal choice? It responds with a quantity read from its reaction function, and this will be less than the amount chosen at the previous stage. By tracing out such a sequence of reactions it is clear that the output of each firm will move to the equilibrium q_E.

Application Box 11.1: Cournot: Fixed costs and brand

Why do we observe so many industries on the national, and even international, stages with only a handful of firms? For example, Intel produces more than half of the world's computer chips, and AMD produces a significant part of the remainder. Why are there only two major commercial aircraft producers in world aviation – Boeing and Airbus? Why are there only a handful of major North American suppliers in pharmaceuticals, automobile tires, soda pop, internet search engines and wireless telecommunications?

The answer lies primarily in the nature of modern product development. Product development (fixed) costs, coupled with a relatively small marginal cost of production, leads to markets where there is enough space for only a few players. The development cost for a new cell phone, or a new aircraft, or a new computer-operating system may run into billions, while the cost of producing each unit may in fact be constant. The enormous development cost associated with many products explains not only why there may be a small number of firms in the domestic market for the product, but also why the number of firms in some sectors is small worldwide.

The Cournot model yields an outcome that lies between monopoly (or collusion/cartel) and competitive market models. It does not necessarily assume that the firms are identical in terms of their cost structure, although the lower-cost producer will end up with a larger share of the market.

The next question that arises is whether this duopoly market will be sustained as a duopoly, or if entry may take place. In particular, if economic profits accrue to the participants will such profits be competed away by the arrival of new producers, or might there be barriers of either a 'natural' or 'constructed' type that operate against new entrants?

11.8 Strategic behaviour: Entry, exit & potential competition

At this point we inquire about the potential entry and impact of new firms – firms who might enter the industry if conditions were sufficiently enticing, meaning the presence of economic profits. One way of examining entry in this oligopolistic world is to envisage potential entry barriers as being either intended or unintended, though the difference between the two can be blurred. Broadly, an unintended or 'natural' barrier is one related to scale economies and the size of the market. An intended barrier involves a strategic decision on the part of the firm to prevent entry.

Unintended entry barriers

Oligopolists tend to have substantial fixed costs, accompanied by declining average costs up to high output levels. Such a cost structure 'naturally' gives rise to a supply side with a small number of suppliers. For examples, given demand and cost structures, could Vancouver support two professional soccer teams; could Calgary support two professional hockey teams; could Montreal sustain two professional football teams? The answer to each of these questions is likely 'no'. Because given the cost structure of these markets, it would not be possible to induce twice as many spectators without reducing the price per game ticket to such a degree that revenue would be insufficient to cover costs. (We will neglect for the moment that the governing bodies of these sports also have the power to limit entry.) Fixed costs include stadium costs, staff payrolls and player payrolls. In fact most costs in these markets are relatively fixed. Market size relative to fixed and variable costs is not large enough to sustain two teams in most cities. Exceptions in reality are huge urban areas such as New York and Los Angeles.

Accordingly, it is possible that the existing team, or teams, may earn economic profit from their present operation; but such profit does not entice further entry, because the market structure is such that the entry of an additional team could lead to each team making losses.

Intended entry barriers

Patent Law

This is one form of protection for incumbent firms. Research and development is required for the development of many products in the modern era. Pharmaceuticals are an example. If innovations were not protected, firms and individuals would not be incentivized to devote their energies and resources to developing new drugs. Society would be poorer as a result. Patent protection is obviously a legal form of protection. At the same time, patent protection can be excessive. If patents provide immunity from replication or copying for an excessive period of time - for longer than required to recoup R & D costs - then social welfare declines because monopoly profits are being generated as a result of output restriction at too high a price.

Advertising

Advertising is a second form of entry deterrence. In this instance firms attempt to market their product as being distinctive and even enviable. For example, *Coca-Cola* and *PepsiCo* invest hundreds of millions annually to project their products in this light. They sponsor sports, artistic and cultural events. Entry into the cola business is not impossible, but brand image is so strong for these firms that potential competitors would have a very low probability of entering this sector profitably. Likewise, in the 'energy-drinks' market, *Red Bull* spends hundreds of millions of dollars per annum on Formula One racing, kite surfing contests, mountain biking events and other extreme sports. In doing this it it reinforcing its brand image and distinguishing its product from Pepsi or Coca-Cola. This form of advertising is one of product differentiation and enables the manufacturer to maintain a higher price for its products by convincing its buyers that there are no close substitutes.

Predatory pricing

This form of pricing constitutes an illegal form of entry deterrence. It involves an incumbent charging an artificially low price for its product in the event of entry of a new competitor. This is done with a view to making it impossible for the entrant to earn a profit. Given that incumbents have generally greater resources than entrants, they can survive a battle of losses for a more prolonged period, thus ultimately driving out the entrant.

An iconic example of predatory pricing is that of Amazon deciding to take on a startup called Quidsi that operated the website *diapers.com*. [3] The latter was proving to be a big hit with consumers in 2009 and Amazon decided that it was eating into Amazon profits on household and baby products. Amazon reacted by cutting its own prices dramatically, to the point where it was ready to loose a huge amount of money in order to grind Quidsi into the ground. The ultimate outcome was that Quidsi capitulated and sold to Amazon.

Whether this was a legal tactic or not we do not know, but it underlines the importance of war chests.

Maintaining a war chest

Many large corporations maintain a mountain of cash. This might seem like an odd thing to do when it could be paying that cash out to owners in the form of dividends. But there are at least two reasons for not doing this. First, personal taxes on dividends are frequently higher than taxes on capital gains; accordingly if a corporation can transform its cash into capital gain by making judicious investments, that strategy ultimately yields a higher post-tax return to the stock holders. A second reason is that a cash war chest serves as a credible threat to competitors of the type described involving Amazon and Quidsi above.

Network externalities

These externalities arise when the existing number of buyers itself influences the total demand for a product. *Facebook* is now a classic example. An individual contemplating joining a social network has an incentive to join one where she has many existing 'friends'. Not everyone views the *Microsoft* operating system (OS) as the best. Many prefer a simpler system such as *Linux* that also happens to be free. However, the fact that almost every new computer (that is not *Apple*) coming onto the market place uses Microsoft OS, there is an incentive for users to continue to use it because it is so easy to find a technician to repair a breakdown.

Transition costs and loyalty cards

Transition costs can be erected by firms who do not wish to lose their customer base. Cell-phone plans are a good example. Contract-termination costs are one obstacle to moving to a new supplier. Some carriers grant special low rates to users communicating with other users within the same network, or offer special rates for a block of users (perhaps within a family). Tim Hortons and other coffee chains offer loyalty cards that give one free cup of coffee for every eight purchased. These suppliers are not furnishing love to their caffeine consumers, they are providing their consumers with an incentive not to switch to a competing supplier. Air miles rewards operate on a similar principle. So too do loyalty cards for hotel chains.

[3] https://slate.com/technology/2013/10/amazon-book-how-jeff-bezos-went-thermonuclear-on-diapers-com.html

How do competitors respond to these loyalty programs? Usually by offering their own. Hilton and Marriot each compete by offering a free night after a given points threshold is reached.

Over-investment

An *over-investment* strategy means that an existing supplier generates additional production capacity through investment in new plant or capital. This is costly to the incumbent and is intended as a signal to any potential entrant that this capacity could be brought on-line immediately should a potential competitor contemplate entry. For example, a ski-resort owner may invest in a new chair-lift, even if she does not use it frequently. The existence of the additional capacity may scare potential entrants. A key component of this strategy is that the incumbent firm invests ahead of time – and inflicts a cost on itself. The incumbent does not simply say "I will build another chair-lift if you decide to develop a nearby mountain into a ski hill." That policy does not carry the same degree of credibility as actually incurring the cost of construction ahead of time. However, such a strategy may not always be feasible: It might be just too costly to pre-empt entry by putting spare capacity in place. Spare capacity is not so different from brand development through advertising; both are types of sunk cost. The threats associated with the incumbent's behaviour become a **credible threat** because the incumbent incurs costs up front.

> **A credible threat** is one that is effective in deterring specific behaviours; a competitor must believe that the threat will be implemented if the competitor behaves in a certain way.

Lobbying

In our chapter on monopoly we stressed the role of political/lobbying activity. Large firms invariably employ public relations firms, and maintain their own public relations departments. The role of these units is not simply to portray a positive image of the corporation to the public; it is to maintain and increase whatever market power such firms already possess. It is as much in the interest of an oligopolistic firm as a monopolist to prevent entry and preserve supernormal profits.

In analyzing perfect competition, we saw that free entry is critical to maintaining normal profits. Lobbying is designed to obstruct entry, and it is also designed to facilitate mergers and acquisitions. The economist Thomas Philippon has written about the increasing concentration of economic power in recent decades in the hands of a small number of corporations in many sectors of the North American economy. He argues that this concentration of power contributes to making the distribution of income more favorable to corporate interests and less favorable to workers. In his recent book ("The Great Reversal: How America Gave up Free Markets"), he shows that, contrary to traditional beliefs, Europe is now much more competitive than the US in most sectors of the economy. More broadband suppliers result in rates in Europe that are about half of US rates. Whereas in the US four airlines control 80% of the market, In Europe they control 40%. If scale economies were the prime determinant of corporate concentration we should not expect such large differences. Likewise, if globalization and technological change were the main determinants of corporate concentration, we should expect experiences in Europe and North America to be similar.

But they are not. Hence, it is reasonable to conclude that entry barriers in North America are more effective, or that regulatory forces are stronger in Europe.

11.9 Matching markets: design

Markets are institutions that facilitate the exchange of goods and services. They act as clearing houses. The normal medium of exchange is money in some form. But many markets deal in exchanges that do not involve money and frequently involve matching: Graduating medical students are normally matched with hospitals in order that graduates complete their residency requirement; in many jurisdictions in the US applicants for places in public schools that form a pool within a given school-board must go through an application process that sorts the applicants into the different schools within the board; patients in need of a new kidney must be matched with kidney donors.

These markets are clearinghouses and have characteristics that distinguish them from traditional currency-based markets that we have considered to this point.

- The good or service being traded is generally *heterogeneous*. For example, patients in search of a kidney donor must be medically compatible with the eventual donor if the organ transplant is not to be rejected. Hospitals may seek residents in particular areas of health, and they must find residents who are, likewise, seeking such placements. Students applying to public schools may be facing a choice between schools that focus upon science or upon the arts. Variety is key.
- Frequently the idea of a market that is mediated by money is *repugnant*. For example, the only economy in the world that permits the sale of human organs is Iran. Elsewhere the idea of a monetary payment for a kidney is unacceptable. A market in which potential suppliers of kidneys registered their reservation prices and demanders registered their willingness to pay is incompatible with our social mores. Consequently, potential living donors or actual deceased donors must be directly matched with a patient in need. While some individuals believe that a market in kidneys would do more good than harm, because a monetary payment might incentivize the availability of many more organs and therefore save many more lives, virtually every society considers the downside to such a trading system to outweigh the benefits.
- Modern matching markets are more frequently *electronically mediated*, and the communications revolution has led to an increase in the efficiency of these markets.

The Economics prize in memory of Alfred Nobel was awarded to Alvin Roth and Lloyd Shapley in 2011 in recognition of their contributions to designing markets that function efficiently in the matching of demanders and suppliers of the goods and services. What do we mean by an efficient mechanism? One way is to define it is similar to how we described the market for apartments in Chapter 5: following an equilibrium in the market, is it possible to improve the wellbeing of one participant without reducing the wellbeing of another? We showed in that example that the market performed efficiently: a different set of renters getting the apartments would reduce total surplus in the system.

Consider a system in which medical graduates are matched with hospitals, and the decision process results in the potential for improvement: Christina obtains a residency in the local University Hospital while Ulrich obtains a residency at the Childrens' Hospital. But Christina would have preferred the Childrens' and Ulrich would have preferred the University. The matching algorithm here was not efficient because, at the end of the allocation process, there is scope for gains for each individual. Alvin Roth devised a matching mechanism that surmounts this type of inefficiency. He called it the deferred acceptance algorithm.

Roth also worked on the matching of kidney donors to individuals in need of a kidney. The fundamental challenge in this area is that a patient in need of a kidney may have a family member, say a sibling, who is willing to donate a kidney, but the siblings are not genetically compatible. The patient's immune system may attack the implantation of a 'foreign' organ. One solution to this incompatibility is to find matching pairs of donors that come from a wider choice set. Two families in each of which there is patient and a donor may be able to cross-donate: donor in Family A can donate to patient in Family B, and donor in Family B can donate to patient in Family A, in the sense that the donor organs will not be rejected by recipients' immune systems. Hence if many patient-donor families register in a clearinghouse, a computer algorithm can search for matching pairs. Surgical operations may be performed simultaneously in order to prevent one donor from backing out following his sibling's receipt of a kidney.

A more recent development concerns 'chains'. In this case a good Samaritan ('unaligned donor') offers a kidney while seeking nothing in return. The algorithm then seeks a match for the good Samaritan's kidney among all of the recipient-donor couples registered in the data bank. Having found (at least) one, the algorithm seeks a recipient for the kidney that will come from the first recipient's donor partner. And so on. It turns out that an algorithm which seeks to maximize the potential number of participating pairs is fraught with technical and ethical challenges: should a young patient, who could benefit from the organ for a whole lifetime, get priority over an older patient, who will benefit for fewer years of life, even if the older patient is in greater danger of dying in the absence of a transplant? This is an ethical problem.

Examples where these algorithms have achieved more than a dozen linked transplants are easy to find on an internet search - they are called *chains*, for the obvious reason.

Consider the following efficiency aspect of the exchange. Suppose a patient has two siblings, each of whom is willing to donate (though only one of the two actually will); should such a patient get priority in the computer algorithm over a patient who has just a single sibling willing to donate? The answer may be yes; the dual donor patient should get priority because if his two siblings have different blood types, this greater variety on the supply side increases the chances for matching in the system as a whole and is therefore beneficial. If a higher priority were not given to the dual-donor patient, there would be an incentive for him to name just one potential donor, and that would impact the efficiency of the whole matching algorithm.

It is not always recognized that the discipline of Economics explores social problems of the nature we have described here, despite the fact that the discipline has developed the analytical tools to address them.

Conclusion

Monopoly and perfect competition are interesting paradigms; but few markets resemble them in the real world. In this chapter we addressed some of the complexities that define the economy we inhabit: It is characterized by strategic planning, entry deterrence, differentiated products and so forth.

Entry and exit are critical to competitive markets. Frequently entry is blocked because of scale economies – an example of a natural or unintended entry barrier. In addition, incumbents can formulate numerous strategies to limit entry.

Firms act strategically – particularly when there are just a few participants in the market. Before acting, firms make conjectures about how their competitors will react, and incorporate such reactions into their own planning. Competition between suppliers can frequently be analyzed in terms of a game, and such games usually have an equilibrium outcome. The Cournot duopoly model that we developed is a game between two competitors in which an equilibrium market output is determined from a pair of reaction functions.

Scale economies are critical. Large development costs or setup costs may mean that the market can generally support just a limited number of producers. In turn this implies that potential new (small-scale) firms cannot benefit from the scale economies and will not survive competition from large-scale suppliers.

Product differentiation is critical. If small differences exist between products produced in markets where there is free entry we get a monopolistically competitive structure. In these markets long-run profits are ‘normal’ and firms operate with some excess capacity. It is not possible to act strategically in this kind of market.

The modern economy also has sectors that have successfully erected barriers. These barriers lead to fewer competitors than could efficiently supply the market. Ultimately the owners of capital are the beneficiaries of these barriers and consumers suffer from higher prices.

Key Terms

Imperfectly competitive firms face a downward-sloping demand curve, and their output price reflects the quantity sold.

Oligopoly defines an industry with a small number of suppliers.

Monopolistic competition defines a market with many sellers of products that have similar characteristics. Monopolistically competitive firms can exert only a small influence on the whole market.

Duopoly defines a market or sector with just two firms.

Concentration ratio: N-firm concentration ratio is the sales share of the largest N firms in that sector of the economy.

Differentiated product is one that differs slightly from other products in the same market.

The **monopolistically competitive equilibrium** in the long run requires the firm's demand curve to be tangent to the ATC curve at the output where $MR = MC$.

Collusion is an explicit or implicit agreement to avoid competition with a view to increasing profit.

Conjecture: a belief that one firm forms about the strategic reaction of another competing firm.

Game: a situation in which contestants plan strategically to maximize their profits, taking account of rivals' behaviour.

Strategy: a game plan describing how a player acts, or moves, in each possible situation.

Nash equilibrium: one in which each player chooses the best strategy, given the strategies chosen by the other player, and there is no incentive for any player to move.

Dominant strategy: a player's best strategy, whatever the strategies adopted by rivals.

Payoff matrix: defines the rewards to each player resulting from particular choices.

Credible threat: one that, after the fact, is still optimal to implement.

Cournot behaviour involves each firm reacting optimally in their choice of output to their competitors' decisions.

Reaction functions define the optimal choice of output conditional upon a rival's output choice.

Exercises for Chapter 11

Exercise 11.1 Imagine that the biggest four firms in each of the sectors listed below produce the amounts defined in each cell. Compute the three-firm and four-firm concentration ratios for each sector, and rank the sectors by degree of industry concentration.

Sector	Firm 1	Firm 2	Firm 3	Firm 4	Total market
Shoes	60	45	20	12	920
Chemicals	120	80	36	24	480
Beer	45	40	3	2	110
Tobacco	206	84	30	5	342

Exercise 11.2 You own a company in a monopolistically competitive market. Your marginal cost of production is \$12 per unit. There are no fixed costs. The demand for your own product is given by the equation $P = 48 - (1/2)Q$.

(a) Plot the demand curve, the marginal revenue curve, and the marginal cost curve.

(b) Compute the profit-maximizing output and price combination.

(c) Compute total revenue and total profit [*Hint*: Remember $AC = MC$ here].

(d) In this monopolistically competitive industry, can these profits continue indefinitely?

Exercise 11.3 Two firms in a particular industry face a market demand curve given by the equation $P = 100 - (1/3)Q$. The marginal cost is \$40 per unit and the marginal revenue is $MR = 100 - (2/3)Q$. The quantity intercepts for demand and MR are 300 and 150.

(a) Draw the demand curve and MR curve to scale on a diagram. Then insert the MC curve.

(b) If these firms got together to form a cartel, what output would they produce and what price would they charge?

(c) Assuming they each produce half of the total what is their individual profit?

Exercise 11.4 The classic game theory problem is the "prisoners' dilemma." In this game, two criminals are apprehended, but the police have only got circumstantial evidence to prosecute them for a small crime, without having the evidence to prosecute them for the major crime of which they are suspected. The interrogators then pose incentives to the crooks-incentives to talk. The crooks are put in separate jail cells and have the option to confess or deny. Their payoff depends upon what course of action each adopts. The payoff matrix is given below. The first element in each box is the payoff (years in jail) to the player in the left column, and the second element is the payoff to the player in the top row.

		B's strategy	
		Confess	Deny
A's strategy	Confess	6,6	0,10
	Deny	10,0	1,1

(a) Does a "dominant strategy" present itself for each or both of the crooks?

(b) What is the Nash equilibrium to this game?

(c) Is the Nash equilibrium unique?

(d) Was it important for the police to place the crooks in separate cells?

Exercise 11.5 Taylormade and Titlelist are considering a production strategy for their new golf drivers. If they each produce a small output, they can price the product higher and make more profit than if they each produce a large output. Their payoff/profit matrix is given below.

		Taylormade strategy	
		Low output	High output
Titleist strategy	Low output	50,50	20,70
	High output	70,20	40,40

(a) Does either player have a dominant strategy here?

(b) What is the Nash equilibrium to the game?

(c) Do you think that a cartel arrangement would be sustainable?

Exercise 11.6 Ronnie's Wraps is the only supplier of sandwich food and makes a healthy profit. It currently charges a high price and makes a profit of six units. However, Flash Salads is considering entering the same market. The payoff matrix below defines the profit outcomes for different possibilities. The first entry in each cell is the payoff/profit to Flash Salads and the second to Ronnie's Wraps.

		Ronnie's Wraps	
		High price	Low price
Flash Salads	Enter the market	2,3	-1,1
	Stay out of market	0,6	0,4

(a) If Ronnie's Wraps threatens to lower its price in response to the entry of a new competitor, should Flash Salads stay away or enter?

(b) Explain the importance of threat credibility here.

Exercise 11.7 *Optional*: Consider the market demand curve for appliances: $P = 3,200 - (1/4)Q$. There are no fixed production costs, and the marginal cost of each appliance is $MC = \$400$. As usual, the MR curve has a slope that is twice as great as the slope of the demand curve.

(a) Illustrate this market geometrically.

(b) Determine the output that will be produced in a 'perfectly competitive' market structure where no profits accrue in equilibrium.

(c) If this market is supplied by a monopolist, illustrate the choice of output.

Exercise 11.8 *Optional*: Consider the outputs you have obtained in Exercise 11.7.

(a) Can you figure out how many firms would produce at the perfectly competitive output? If not, can you think of a reason?

(b) If, in contrast, each firm in that market had to cover some fixed costs, in addition to the variable costs defined by the MC value, would that put a limit on the number of firms that could produce in this market?

Part Five

The Factors of Production

Output is produced by the factors of production – labour, capital, and land. In Chapter 12 we explore how the markets for these factors, or productive inputs, operate. We focus primarily on labour and capital.

Chapter 13 deals with human capital – a broader concept than labour. The human capital embodied in a person depends upon her skills, education and experience, and this combination determines the rewards, or return, she obtains when employed. This chapter also explores the distribution of income in Canada and the degree of inequality in that distribution. It concludes with a brief overview of the 'top one percent'.

Chapter 12

Labour and capital

In this chapter we will explore:

The chapter deals with the markets for the factors of production—labour and capital. The analysis will be presented in terms of the demand, supply and market equilibrium for each. While this is a standard analytical approach in microeconomics, the markets for labour and capital differ from goods and services markets.

In the first instance, goods and services are purchased and consumed by the buyers. In contrast, labour and capital are used as inputs in producing those 'final' goods and services. So the value of labour and capital to a producer depends in part upon the value of the products that the labour and capital are used to produce. Economists say that the value of the factors of production *derives* from the value of the products they ultimately produce.

Secondly, labour and capital offer services. When an employer hires a worker, that worker supplies her time and skills and energy to the employer. When a piece of equipment is rented to a producer, or purchased by a producer, that equipment provides a stream of productive services also. The employer does not purchase the worker, given that we do not live in a society where slavery is legal. In contrast, she may decide to purchase the capital, or else rent it.

The third characteristic of these markets is the time dimension associated with labour and capital. Specifically, once built, a machine will customarily have a lifetime of several years, during which it depreciates in value. Furthermore, it may become obsolete on account of technological change before the end of its anticipated life. Labour too may become obsolete, or at least lose some of its value with the passage of time, if the skills embodied in the labour cease to be required in the economy.

12.1 Labour – a derived demand

The value of labour springs from the value of its use, that is the value placed upon goods and services that it produces – product prices. The wage is the price that equilibrates the supply and demand for a given type of labour, and it reflects the value of that labour in production. Formally, the demand for labour (and capital) is thus a **derived demand**, in contrast to being a 'final' demand.

> **Demand for labour**: a derived demand, reflecting the value of the output it produces.

We must distinguish between the long run and the short run in our analysis of factor markets. On the *supply side* certain factors of production are fixed in the short run. For example, the supply of radiologists can be increased only over a period of years. While one hospital may be able to attract radiologists from another hospital to meet a shortage, this does not increase the supply in the economy as a whole.

On the *demand side* there is the conventional difference between the short and long run: In the short run some of a firm's factors of production, such as capital, are fixed, and therefore the demand for labour differs from when all factors are variable – the long run.

Demand in the short run

Table 12.1 contains information from the example developed in Chapter 8. It can be used to illustrate how a firm reacts in the short run to a change in an input price, or to a change in the output price. The response of a producer to a change in the wage rate constitutes a demand function for labour – a schedule relating the quantity of the input demanded to different input prices. The output produced by the various numbers of workers yields a marginal product curve, whose values are stated in column 3. The marginal product of labour, MP_L, as developed in Chapter 8, is the additional output resulting from one more worker being employed, while holding constant the other (fixed) factors. But what is the *dollar value* to the firm of an additional worker? It is the additional *value of output* resulting from the additional employee – the price of the output times the worker's marginal contribution to output, his MP. We term this the **value of the marginal product**.

> The **value of the marginal product** is the marginal product multiplied by the price of the good produced.

Table 12.1: Short-run production and labour demand

Workers (1)	**Output** (2)	MP_L (3)	$VMP_L = MP_L \times P$ (4)	**Marginal profit =** (VMP_L**−wage**) (5)
0	0			
1	15	15	1050	150
2	40	25	1750	750
3	70	30	2100	1100
4	110	40	2800	1800
5	145	35	2450	1450
6	175	30	2100	1100
7	200	25	1750	750
8	220	20	1400	400
9	235	15	1050	50
10	240	5	350	negative

Each unit of labour costs $1,000; output sells at a fixed price of $70 per unit.

In this example the MP_L first rises as more labour is employed, and then falls. With each unit of output selling for $70 the value of the marginal product of labour (VMP_L) is given in column 4. The first worker produces 15 units each week, and since each unit sells for a price of $70, his production value to the firm is $1,050 ($= \$70 \times 15$). A second worker produces 25 units, so his value to the firm is $1,750, and so forth. If the weekly wage of each worker is $1,000 then the firm can estimate its marginal profit from hiring each additional worker. This is the difference between the value of the marginal product and the wage paid, and is given in the final column of the table.

It is profitable to hire more workers as long as the cost of an extra worker is less than the VMP_L. The equilibrium amount of labour to employ is therefore 9 units in this example. If the firm were to hire one more worker the contribution of that worker to its profit would be negative ($\$350 - \$1,000$), and if it hired one worker less it would forego the opportunity to make an additional profit of $50 on the 9th unit ($\$1,050 - \$1,000$).

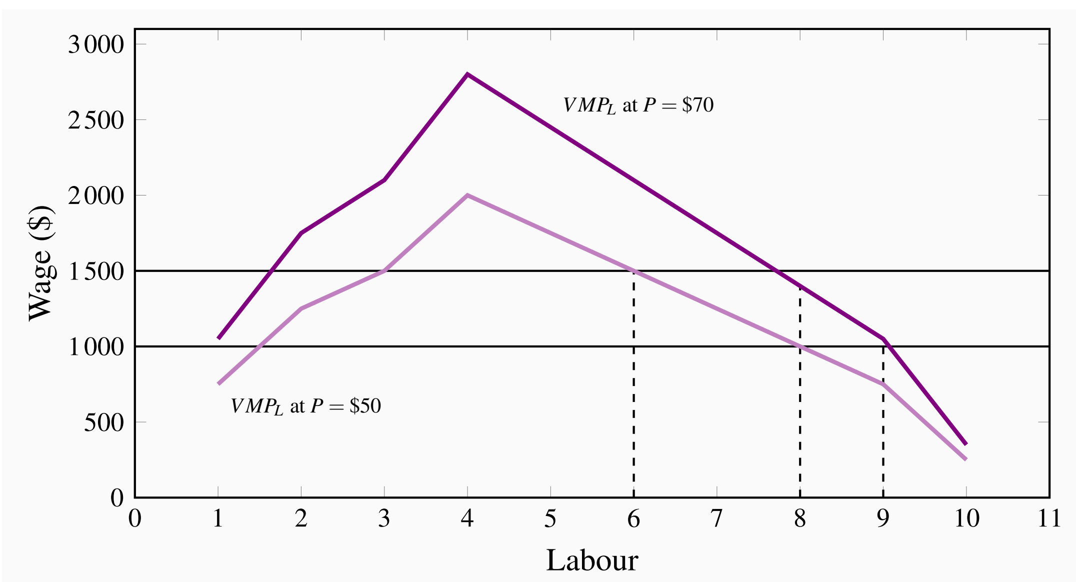
Profit maximizing hiring rule:

- *If the VMP_L of next worker $>$ wage, hire more labour.*
- *If the VMP_L $<$ wage, hire less labour.*

To this point we have determined the profit maximizing amount of labour to employ when the output price and the wage are given. However, a demand function for labour reflects the demand for labour at many different wage rates, just as the demand function for any product reflects the quantity demanded at various prices. Accordingly, suppose the wage rate is \$1,500 per week rather than \$1,000. The optimal amount of labour to employ in this case is determined in exactly the same manner: Employ the amount of labour where its contribution is marginally profitable. Clearly the optimal amount to employ is 7 units: The value of the seventh worker to the firm is \$1,750 and the value of the eighth worker is \$1,400. Hence it would not be profitable to employ the eighth, because his marginal contribution to profit would be negative. Following the same procedure we could determine the optimal amount of labour to employ at different wages. This implies that the VMP_L function is the demand for labour function because it determines the most profitable amount of labour to employ at any wage.

The optimal amount of labour to hire is illustrated in Figure 12.1. The wage and VMP_L curves come from Table 12.1. The VMP_L curve has an upward sloping segment, reflecting increasing productivity, and then a regular downward slope as developed in Chapter 8. At employment levels where the VMP_L is greater than the wage additional labour should be employed. But when the VMP_L falls below the wage rate employment should stop. If labour is divisible into very small units, the optimal employment decision is where the MP_L function intersects the wage line.

Figure 12.1: The demand for labour



The optimal hiring decision is defined by the condition that the value of the MP_L is greater than or equal to the wage paid. 9 workers are employed when the wage is \$1,000 and the price of output is \$70; 6 workers are employed when the wage is \$1,500 and the price of output is \$50.

Figure 12.1 also illustrates what happens to hiring when the output price changes. Consider a reduction in its price to \$50 from \$70. The profit impact of such a change is negative because the value of each worker's output has declined. Accordingly, the demand curve must reflect this by shifting inward (down), as in the figure. At various wage rates, less labour is now demanded. The new VMP_L schedule can be derived in Table 12.1 as before: It is the MP_L schedule multiplied by the lower value ($50) of the final good.

In this example the firm is a perfect competitor in the output market, because the price of the good it produces is fixed. It can produce and sell more of the good without this having an impact on the price of the good in the marketplace. Where the firm is not a perfect competitor it faces a declining MR function. In this case the value of the MP_L is the product of MR and MP_L rather than P and MP_L. To distinguish the different output markets we use the term **marginal *revenue* product of labour** (MRP_L) when the demand for the output slopes downward. But the optimizing principle remains the same: The firm should calculate the value of each additional unit of labour, and hire up to the point where the additional revenue produced by the worker exceeds or equals the additional cost of that worker.

> The **marginal revenue product of labour** is the additional revenue generated by hiring one more unit of labour where the marginal revenue declines.

Demand in the long run

In Chapter 8 we proposed that firms choose their factors of production in accordance with cost-minimizing principles. In producing a specific output, firms choose the least-cost combination of labour and plant size. But how is this choice affected when the price of labour or capital changes? While adjustment to price changes may require a long period of time, we know that if one factor becomes more (less) expensive, the firm will likely change the mix of capital and labour away from (towards) that factor. For example, when the accuracy and prices of production robots began to fall in the nineteen nineties, auto assemblers reduced their labour and used robots instead. When computers and computer software improved and declined in price, clerical workers were replaced by computers that were operated by accountants. But such adjustments and responses do not occur overnight.

In the short run a higher wage increases costs, but the firm is constrained in its choice of inputs by a fixed plant size. In the long run, a wage increase will induce the firm to use relatively more capital than when labour was less expensive in producing a given output. But despite the new choice of inputs, a rise in the cost of any input must increase the total cost of producing any output.

A change in the price of any factor has two impacts on firms: In the first place producers will *substitute* away from the factor whose price increases; second, there will be an impact on output and a change in the price of the final good it produces. Since the cost structure increases when the price of an input rises, the supply curve in the market for the good must reflect this – any given output will now be supplied at a higher price. With a downward sloping demand, this shift in supply must increase the price of the good and reduce the amount sold. This second effect can be called an *output effect*.

Monopsony

Some firms may have to pay a higher wage in order to employ more workers. Think of Hydro Quebec building a dam in Northern Quebec. Not every hydraulic engineer would be equally happy working there as in Montreal. Some engineers may demand only a small wage premium to work in the North, but others will demand a high premium. If so, Hydro Quebec must pay a higher wage to attract more workers – it faces an upward sloping supply of labour curve. Hydro Quebec is the sole buyer in this particular market and is called a **monopsonist** – a single buyer. Our general optimizing principle governing the employment of labour still holds, even if we have different names for the various functions: Hire any factor of production up to the point where the cost of an additional unit equals the value generated for the firm by that extra worker. The essential difference here is that when a firm faces an upward sloping labour supply it will have to pay more to attract additional workers and *also pay more to its existing workers*. This will impact the firm's willingness to hire additional workers.

> A **monopsonist** is the sole buyer of a good or service and faces an upward-sloping supply curve.

Application Box 12.1: Monopsonies

Monopsonies are more than a curiosity; they exist in the real world. An excellent example is the cannabis market in Canada. Virtually every province has set up a trading agency that has the sole right to purchase cannabis from growers; growers and processors are not permitted to sell directly to retailers; they may only sell to the monopsony by law. In turn, these provincial cannabis monopsonies are frequently retail monopolists in that the agency owns all of the retail outlets in the province.

Firm versus industry demand

The demand for labour within an industry, or sector of the economy, is obtained from the sum of the demands by each individual firm. It is analogous to the goods market, but with a subtle difference. In Chapter 3 we obtained a market demand by summing individual demands horizontally. The same could be done here: At lower (or higher) wages, each firm will demand more (or less) labour. However, if all firms employ more labour in order to increase their output, the price of the output will likely decline. This in turn will moderate the demand for labour – it is slightly less valuable now that the price of the output it produces has fallen. This is a subtle point, and we can reasonably think of the demand for labour in a given sector of the economy as the sum of the demands on the part of the employers in that sector.

12.2 The supply of labour

Most prime-age individuals work, but some do not. The decision to join the **labour force** is called the **participation decision**. Of those who do participate in the labour force, some individuals work full time, others work part time, and yet others cannot find a job. The **unemployment rate** is the fraction of the labour force actively seeking employment that is not employed.

> The **participation rate** for the economy is the fraction of the population in the working age group that joins the labour force.
>
> The **labour force** is that part of the population either employed or seeking employment.
>
> The **unemployment rate** is the fraction of the labour force actively seeking employment that is not employed.

Data on participation rates in Canada are given in Table 12.2 below for specific years in the modern era. The overall participation for men and women combined has increased since 1977 from 60.8% to 65.8% This aggregated rate camouflages different patterns for men and women. The rates for women have been rising while the rates for men have fallen. Women today are more highly educated, and their role in society and the economy is viewed very differently than in the earlier period. Female participation has increased both because of changing social norms, a rise in household productivity, the development of service industries designed to support home life, and the development of the institution of daycare for young children.

In contrast, male participation rates declined over the period, largely offsetting the increase in female participation. Fewer individuals in total are retiring before the age of 55 in the most recent decades. This reflects both the greater number of females in the market place, and perhaps also a recognition that many households have not saved enough to fund a retirement period that has become longer as a result of increased longevity.

Table 12.2: Labour force participation rate, Canada 1977-2015

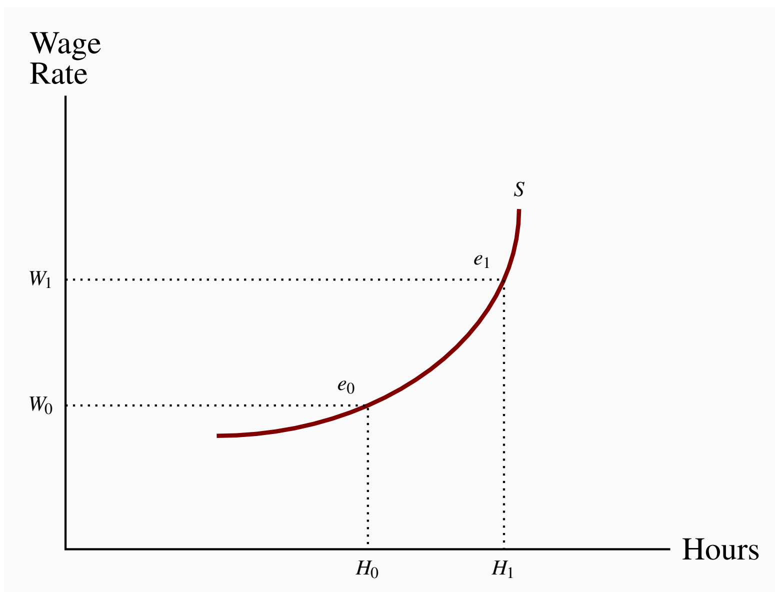
Year	Total	Men	Women	All > 55	Unemployment
1977	60.8	80.2	42.1	30.5	5.9
1990	66.6	77.1	56.8	25.9	7.4
1994	65.4	74.9	56.8	24.5	9.2
2001	66.1	73.4	59.2	26.4	6.2
2008	67.5	73.6	61.6	34.2	5.1
2015	66.2	72.0	60.6	37.3	6.0
2019	65.8	71.0	60.8	38.0	4.4

Source: Statistics Canada, CANSIM 14-10-0287-02
September of each year, for individuals aged ≥ 25, unless stated.

At the micro level, the participation rate of individuals depends upon several factors. First, the wage rate that an individual can earn in the market is crucial. If that wage is low, then the individual may be more efficient in producing home services directly, rather than going into the labour market, earning a modest income and having to pay for home services. Second, there are fixed costs associated with working. A decision to work means that the individual must have work clothing, must undertake the costs of travel to work, and pay for daycare if there are children in the family. Third, the participation decision depends upon non-labour income. If the individual in question has a partner who earns a substantial amount, or if she has investment income, she will have less incentive to participate. Fourth, it depends inversely upon the tax rate.

The supply curve relates the supply decision to the price of labour – the wage rate. Economists who have studied the labour market tell us that the individual supply curve is upward sloping: As the wage increases, the individual wishes to supply more labour. From the point e_0 on the supply function in Figure 12.2, let the wage increase from W_0 to W_1.

Figure 12.2: Individual labour supply



A wage increase from W_0 to W_1 induces the individual to *substitute* away from leisure, which is now more expensive, and work *more*. But the higher wage also means the individual can work fewer hours for a given standard of living; therefore the income effect induces *fewer hours*. On balance the substitution effect tends to dominate and the supply curve therefore slopes upward.

The individual offers more labour, H_1, at the higher wage. What is the economic intuition behind the higher amount of labour supplied? Like much of choice theory there are two impacts associated with a higher price. First, the higher wage makes leisure more expensive relative to working. That midweek game of golf has become more expensive in terms of what the individual could earn. So the individual should *substitute* away from the more expensive 'good', leisure, towards labour. But at the same time, in order to generate a given income target the individual can work fewer hours at the higher wage. This is a type of *income effect*, indicating that income is greater at a higher wage regardless of the amount worked, and this induces the individual to work less. The fact that we draw the labour supply curve with a positive slope means that the substitution effect is the more important of the two. That is what statistical research has revealed.

Elasticity of the supply of labour

The value of the supply elasticity depends upon how the market in question is defined. In particular, it depends upon how large or small a given sector of the economy is, and whether we are considering the short run or the long run.

Suppose an industry is small relative to the whole economy and employs workers with common skills. These industries tend to pay the 'going wage'. For example, very many students are willing to work at the going rate for telemarketing firms, which compose a small sector of the economy. This means that the supply curve of such labour, *as far as that sector is concerned*, is in effect horizontal – infinitely elastic.

But some industries may not be small relative to the total labour supply. And in order to get more labour to work in such large sectors it may be necessary to provide the inducement of a higher wage: Additional workers may have to be attracted from another sector by means of higher wages. To illustrate: Consider the behaviour of two related sectors in housing – new construction and home restoration. In order to employ more plumbers and carpenters, new home builders may have to offer higher wages to induce them to move from the renovation sector. In this case the new housing industry's labour supply curve slopes upwards.

In the time dimension, a longer period is always associated with more flexibility. In this context, the supply of labour to any sector is more elastic, because it may take time for workers to move from one sector to another. Or, in cases where skills must be built up: When a sectoral expansion bids up the wages of information technology (IT) workers, more school leavers are likely to develop IT skills. Time will be required before additional graduates are produced, but in the long run, such additional supply will moderate the short-run wage increases.

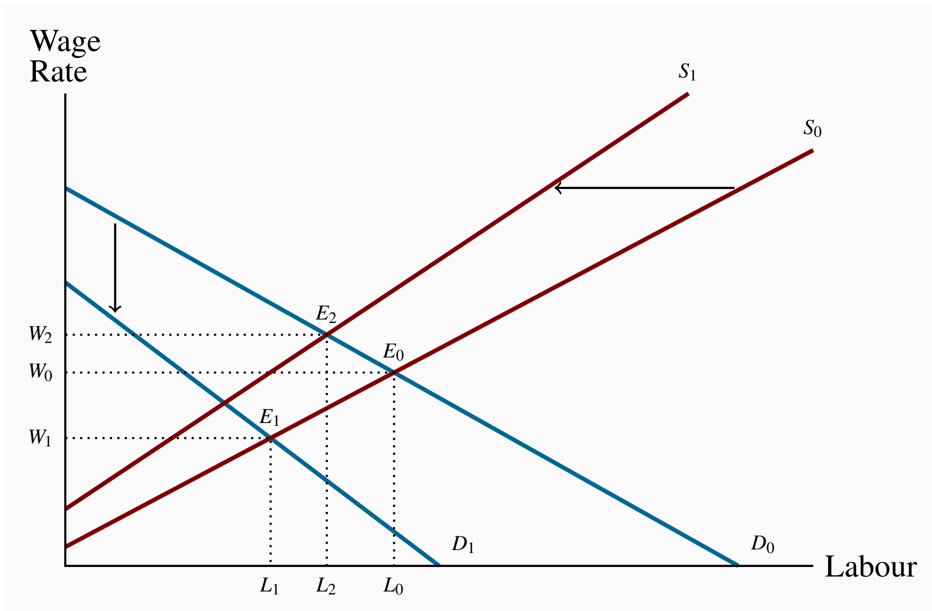
Wages can be defined as being before-tax or after-tax. The after-tax, or take-home, wage is more important than the gross wage in determining the quantity of labour to be supplied. If taxes on additional hours of work are very high, workers are more likely to supply less hours than if tax rates are lower.

12.3 Labour market equilibrium and mobility

The fact that labour is a derived demand differentiates the labour market's equilibrium from the goods-market equilibrium. Let us investigate this with the help of Figure 12.3; it contains supply and demand functions for one particular industry – the cement industry, let us assume.

In Figure 12.1 we illustrated the impact on the demand for labour of a decline in the price of the output produced – a decline in the output price reduced the value of the marginal product of labour. In the current example, suppose that a slowdown in construction results in a decline in the price of cement. The impact of this price fall is to reduce the output value of each worker in the cement producing industry, because their output now yields a lower price. This decline in the VMP_L is represented in Figure 12.3 as a shift from D_0 to D_1, which results in the new equilibrium E_1.

Figure 12.3: Equilibrium in an industry labour market



A fall in the *price of the good* produced in a particular industry reduces the value of the MP_L. Demand for labour thus falls from D_0 to D_1 and a new equilibrium E_1 results. Alternatively, from E_0, an increase in wages in another sector of the economy induces some labour to move to that sector. This is represented by the shift of S_0 to S_1 and the new equilibrium E_2.

As a second example: Suppose that wages in some other sectors of the economy increase. The impact of this on the cement sector is that the supply of labour to the cement sector is reduced. In Chapter 3 we showed that a change in other prices may *shift* the demand or supply curve of interest. In Figure 12.3 supply shifts from S_0 to S_1 and the equilibrium goes from E_0 to E_2.

How large are these impacts likely to be? That will depend upon how mobile labour is between sectors: Spillover effects will be smaller if labour is less mobile. This brings us naturally to the concepts of **transfer earnings** and **rent**.

Transfer earnings and rent

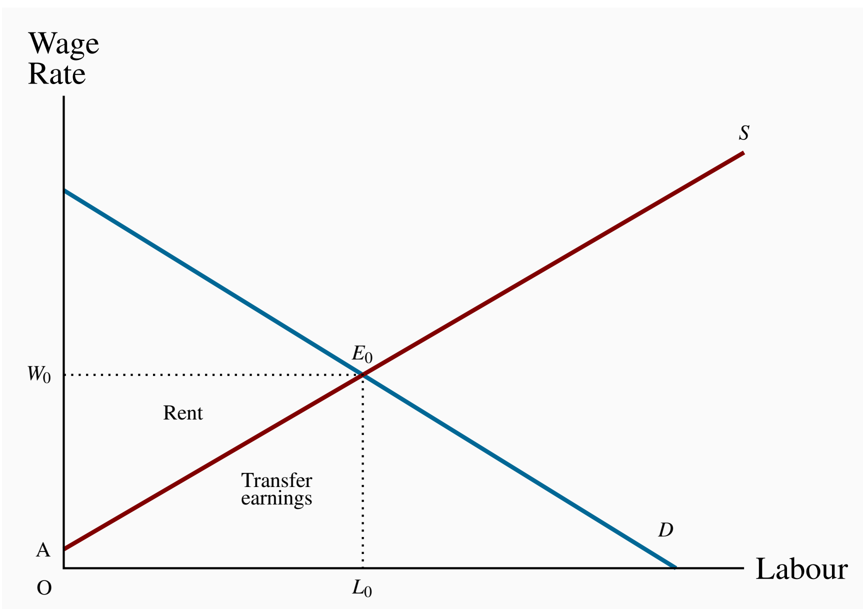
Consider the case of a performing violinist whose wage is $80,000. If, as a best alternative, she can earn $60,000 as a music teacher then her rent is $20,000 and her transfer earnings $60,000:

Her rent is the excess she currently earns above the best alternative. Another violinist in the same orchestra, earning the same amount, who could earn \$55,000 as a teacher has rent of \$25,000. The alternative is also called the **reservation wage**. The violinists should not work in the orchestra unless they earn at least what they can earn in the next best alternative.

> **Transfer earnings** are the amount that an individual can earn in the next highest paying alternative job.
>
> **Rent** is the excess remuneration an individual currently receives above the next best alternative. This alternative is the **reservation wage**.

These concepts are illustrated in Figure 12.4. In this illustration, different individuals are willing to work for different amounts, but all are paid the same wage W_0. The market labour supply curve by definition defines the wage for which each individual is willing to work. Thus the rent earned by labour in this market is the sum of the excess of the wage over each individual's transfer earnings – the area W_0E_0A. This area is also what we called producer or supplier surplus in Chapter 5.

Figure 12.4: Transfer earnings and rent



Rent is the excess of earnings over reservation wages. Each individual earns W_0 and is willing to work for the amount defined by the labour supply curve. Hence rent is W_0E_0A and transfer earnings OAE_0L_0. Rent is thus the term for supplier surplus in this market.

Free labour markets?

Real-world labour markets are characterized by trade unions, minimum wage laws, benefit regulations, severance packages, parental leave, sick-day allowances and so forth. So can we really claim that markets work in the way we have described them – essentially as involving individual agents demanding and supplying labour? While labour markets are not completely 'free' in the conventional sense, the important issue is whether these interventions, that are largely designed to protect workers, have a large or small impact on the market. One reason why unemployment rates are generally higher in European economies than in Canada and the US is that labour markets are less subject to controls, and workers have a less supportive social safety net in North America.

Application Box 12.2: Are high salaries killing professional sports?

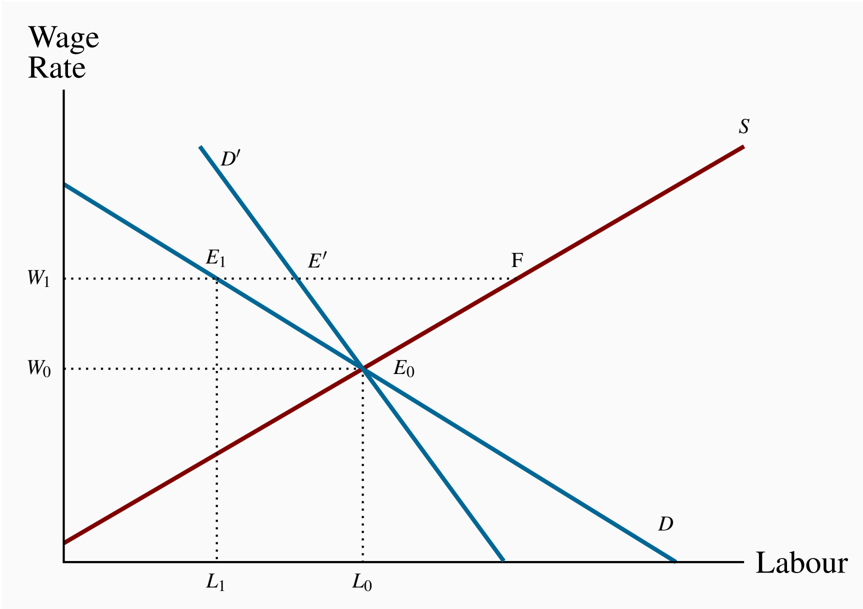
It is often said that the agents of professional players are killing their sport by demanding unreasonable salaries. On occasion, the major leagues are threatened with strikes, even though players are paid millions each year. In fact, wages are high because the derived demand is high. Fans are willing to pay high ticket prices, and television rights generate huge revenues. Combined, these revenues not only make ownership profitable, but increase the demand for the top players.

The lay person may be horrified at thirty-million dollar annual salaries. But in reality, many players receiving such salaries may be earning less than their marginal product! If Tom Brady did not play for the New England Patriots the team would have a lower winning record, attract fewer fans and make less profit. If Brady is paid \$25m per season, he is being paid less than his marginal product if the team were to lose \$40m in revenue as a result of his absence.

Given this, why do some teams incur financial losses? In fact very few teams make losses: Cries of poverty on the part of owners are more frequently part of the bargaining process, and revenue sharing means that very few teams do not make a profit.

The impact of 'frictions', such as unionization and minimum wages, in the labour market can be understood with the help of Figure 12.5. The initial 'free market' equilibrium is at E_0, assuming that the workers are not unionized. In contrast, if the workers in this industry form a union, and negotiate a higher wage, for example W_1 rather than W_0, then fewer workers will be employed. But how big will this reduction be? Clearly it depends on the elasticities of demand and supply. With the demand curve D, the excess supply at the wage W_1 is the difference E_1F. However, if the demand curve is less elastic, as illustrated by the curve D', the excess supply is $E'F$. The excess supply also depends upon the supply elasticity. It is straightforward to see that a less elastic (more vertical) supply curve through E_0 would result is less excess supply.

Figure 12.5: Market interventions



E_0 is the equilibrium in the absence of a union. If the presence of a union forces the wage to W_1 fewer workers are employed. The magnitude of the decline from L_0 to L_1 depends on the elasticity of demand for labour. The excess supply at the wage W_1 is (F-E_1). With a less elastic demand curve (D') the excess supply is reduced to (F-E').

Beyond elasticity, the magnitude of the excess supply will also depend upon the degree to which the minimum wage, or the union-negotiated wage, lies above the equilibrium. That is, a larger value of the difference $(W_1 - W_0)$ results in more excess supply than a smaller difference.

While the above discussion pertains to unionization, it could equally well be interpreted in a minimum-wage context. If this figure describes the market for low-skill labour, and the government intervenes by setting a legal minimum at W_1, then this will induce some degree of excess supply, depending upon the actual value of W_1 and the elasticities of supply and demand.

Despite the fact that a higher wage may induce some excess supply, it may increase total earnings. In Chapter 4 we saw that the dollar value of expenditure on a good increases when the price rises if the demand is inelastic. In the current example the 'good' is labour. Hence, a union-negotiated wage increase, or a higher minimum wage will each increase total remuneration if the demand for labour is inelastic. A case which has stirred great interest is described in Application Box 12.3.

Application Box 12.3: David Card on minimum wage

David Card is a famous Canadian-born labour economist who has worked at Princeton University and University of California, Berkeley. He is a winner of the prestigious Clark medal, an award made annually to an outstanding economist under the age of forty. Among his many contributions to the discipline, is a study of the impact of minimum wage laws on the employment of fast-food workers. With Alan Krueger as his co-researcher, Card examined the impact of the 1992 increase in the minimum wage in New Jersey and contrasted the impact on employment changes with neighbouring Pennsylvania, which did not experience an increase. They found virtually no difference in employment patterns between the two states. This research generated so much interest that it led to a special conference. Most economists now believe that modest changes in the level of the minimum wage have a small impact on employment levels.

Since about 2015, numerous labor-friendly movements favoring higher wages for low-paid workers have proposed a $15 minimum in both Canada and the US. Some political parties have supported this movement, as have specific cities and municipalities and governments. While any increase in the minimum wage must by definition help those working, care must be exercised in implementing particularly large increases. This is because large increases in particular areas or spheres may induce production units to move outside of the area covered, and thereby shift jobs to lower-wage areas.

12.4 Capital – concepts

The share of national income accruing to capital is more substantial than commonly recognized. National income in Canada is divided 60-40, favoring labour. This leaves a very large component going to the owners of capital. The stock of **physical capital** includes assembly-line machinery, rail lines, dwellings, consumer durables, school buildings and so forth. It is the stock of produced goods used as inputs to the production of other goods and services.

> **Physical capital** is the stock of produced goods that are inputs in the production of other goods and services.

Physical capital is distinct from land in that the former is produced, whereas land is not. These in turn differ from *financial wealth*, which is not an input to production. We add to the capital stock by undertaking investment. But, because capital depreciates, investment in new capital goods is required merely to stand still. **Depreciation** accounts for the difference between **gross and net investment**.

> **Gross investment** is the production of new capital goods and the improvement of existing capital goods.
>
> **Net investment** is gross investment minus depreciation of the existing capital stock.
>
> **Depreciation** is the annual change in the value of a physical asset.

Since capital is a **stock** of productive assets we must distinguish between the value of services that **flow** from capital and the value of capital assets themselves.

> A **stock** is the quantity of an asset at a point in time.
>
> A **flow** is the stream of services an asset provides during a period of time.

When a car is rented it provides the driver with a service; the car is the asset, or stock of capital, and the driving, or ability to move from place to place, is the service that flows from the use of the asset. When a photocopier is leased it provides a stream of services to the user. The copier is the asset; it represents a stock of physical capital. The printed products result from the service the copier provides per unit of time.

The **price of an asset** is what a purchaser pays for the asset. The owner then obtains the future stream of capital services it provides. Buying a car for $30,000 entitles the owner to a stream of future transport services. The term **rental rate** defines the cost of the services from capital.

> **Capital services** are the production inputs generated by capital assets.
>
> The **rental rate** is the cost of using capital services.
>
> The **price of an asset** is the financial sum for which the asset can be purchased.

But what determines the *price* of a productive asset? The price must reflect the value of future services that the capital provides. But we cannot simply add up these future values, because a dollar today is more valuable than a dollar several years from now. The key to valuing an asset lies in understanding how to compute the *present value* of a future income stream.

Present values and discounting

When capital is purchased it generates a stream of dollar values (returns) in the future. A critical question is: How is the price that should be paid for capital today related to the benefits that capital will bring in the future? Consider the simplest of examples: A business is contemplating buying a computer. This business has a two-year horizon. It believes that the purchase of the computer will yield a return of $500 in the first year (today), $500 in the second year (one period into the future), and have a scrap value of \$200. What is the maximum price the entrepreneur should pay for the computer? The answer is obtained by *discounting* the future returns to the present. Since a dollar today is worth more than a dollar tomorrow, we cannot simply add the dollar values from different time periods.

The value today of \$500 received a year from now is less than \$500, because if you had this amount today you could invest it at the going rate of interest and end up with more than \$500 tomorrow. For example, if the rate of interest is 10% $(= 0.1)$, then \$500 today is worth \$550 next period. By the same reasoning, \$500 tomorrow is worth less than \$500 today. Formally, the value next period of any amount is that amount plus the interest earned; in this case the value next period of \$500 today is $\$500 \times (1+r) = \$500 \times 1.1 = \$550$, where r is the interest rate. It follows that if we multiply a given sum by $(1+r)$ to obtain its value next period, then we must divide a sum received next period to obtain its value today. Hence the value today of \$500 next period is simply $\$500/(1+r) = \$500/1.1 = \$454.54$. To see that this must be true, note that if you have \$454.54 today you can invest it and obtain \$500 next period if the interest rate is 10%. In general:

$$\text{Value next period} = \text{value this period} \times (1 + \text{interest rate})$$

$$\text{Value this period} = \frac{\text{Value next period}}{(1 + \text{interest rate})}.$$

This rule carries over to any number of future periods. The value of a sum of money today two periods into the future is obtained by multiplying the today value by $(1 + \text{interest rate})$ twice. Or the value of a sum of money today that will be received two periods from now is that sum divided by $(1 + \text{interest rate})$ twice. And so on, for any number of time periods. So if the amount is received twenty years into the future, its value today would be obtained by dividing that sum by $(1 + \text{interest rate})$ twenty times; if received 'n' periods into the future it must be divided by $(1 + \text{interest rate})$ 'n' times.

Two features of this discounting are to be noted: First, if the interest rate is high, the value today of future sums is smaller than if the interest rate is low. Second, sums received far in the future are worth much less than sums received in the near future.

Let us return to our initial example, assuming the interest rate is 0.1 (or 10%). The value of the year 1 return is \$500. The value of the year 2 return today is \$454.54, and the scrap value in today's terms is \$181.81. The value of all returns discounted to today is thus \$1,136.35.

Table 12.3: Present value of an asset ($i = 10\%$)

Year	Annual return	Scrap value	Discounted values
Year 1	500		500
Year 2	500	200	454.54 + 181.81
Asset value today			1,136.35

> The **present value of a stream of future earnings** is the sum of each year's earnings divided by one plus the interest rate 'n' times, where 'n' is the number of years in the future when the amount will be received.

We are now in a position to determine how much the buyer should be willing to pay for the computer. Clearly if the value of the computer today, measured in terms of future returns to the entrepreneur's business, is \$1,136.35, then the potential buyer should be willing to pay any sum less than that amount. Paying more makes no economic sense.

Discounting is a technique used in countless applications. It underlies the prices we are willing to pay for corporate stocks: Analysts make estimates of future earnings of corporations; they then *discount those earnings back to the present*, and suggest that we not pay more for a unit of stock than indicated by the present value of future earnings.

12.5 The capital market

Demand

The analysis of the demand for the services of capital parallels closely that of labour demand: The rental rate for capital replaces the wage rate and capital services replace the hours of labour. It is important to keep in mind the distinction we drew above between capital services on the one hand and the amount of capital on the other. Capital services are produced by capital assets, just as work is produced by humans. Terms that are analogous to the marginal product of labour emerge naturally: The **marginal product of capital** (MP_K) is the output produced by one additional unit of capital services, with other inputs held constant. The **value of this marginal product** (VMP_K) is its value in the market place. It is the MP_K multiplied by the price of output.

The MP_K must eventually decline with a fixed amount of other factors of production. So, if the price of output is fixed for the firm, it follows that the VMP_K must also decline. We could pursue an analysis of the short-run demand for capital services, assuming labour was fixed, that would completely mirror the short-run demand for labour that we have already developed. But this would not add any new insights, so we move on to the supply side.

> The **marginal product of capital** is the output produced by one additional unit of capital services, with all other inputs being held constant.
>
> The **value of the marginal product of capital** is the marginal product of capital multiplied by the price of the output it produces.

Supply

We can grasp the key features of the market for capital by recognizing that the *flow* of capital services is determined by the capital *stock*: More capital means more services. The analysis of supply is complex because we must distinguish between the long run and the short run, and also between the supply to an industry and the supply in the whole economy.

In the *short run* the total supply of capital assets, and therefore services, is fixed to the *economy*, since new production capacity cannot come on stream overnight: The short-run supply of services is therefore vertical. In contrast, *a particular industry* in the short run faces a positively sloped supply: By offering a higher rental rate for trucks, one industry can bid them away from others.

The *long run* is a period of sufficient length to permit an addition to the capital stock. A supplier of capital, or capital services, must estimate the likely return he will get on the equipment he is contemplating having built. To illustrate: He is analyzing the purchase or construction of an earthmover that will cost \$100,000. Assuming that the annual maintenance and depreciation costs are \$10,000, and that the interest rate is 5% (implying that annual interest cost is \$5,000), it follows that the annual cost of owning such a machine is \$15,000. If the entrepreneur is to undertake the investment she must therefore earn at least this amount annually (by renting it to others, or using it herself), and this is what is termed the **required rental**. We can think of it as the opportunity cost of ownership.

The **required rental** covers the sum of *maintenance*, *depreciation* and *interest costs*.

Prices and returns

In the long run, capital services in any sector of the economy must earn the required rental. If they earn more, entrepreneurs will be induced to build or purchase additional capital goods; if they earn less, owners of capital will allow machines to depreciate, or move the machines to other sectors of the economy.

As an example, the price of oil on world markets fell by half during 2015; from about \$100US per barrel to \$50US. At this price, many oil wells were no longer profitable, and oil drilling equipment was decommissioned. Technically, the value of the marginal product of capital declined, because the price of the good it was producing declined. In the near and medium term, no new investment in capital goods will take place in the oil drilling sector of the economy. If the price of oil should increase in the future, some of the decommissioned capital will be brought back into service. But some of this capital will deteriorate or depreciate and simply 'die', and be sold for scrap metal – particularly the older vintage capital. Only when the stock of oil drilling equipment is reduced by depreciation and decay to the required level will any new investment in this form of capital take place.

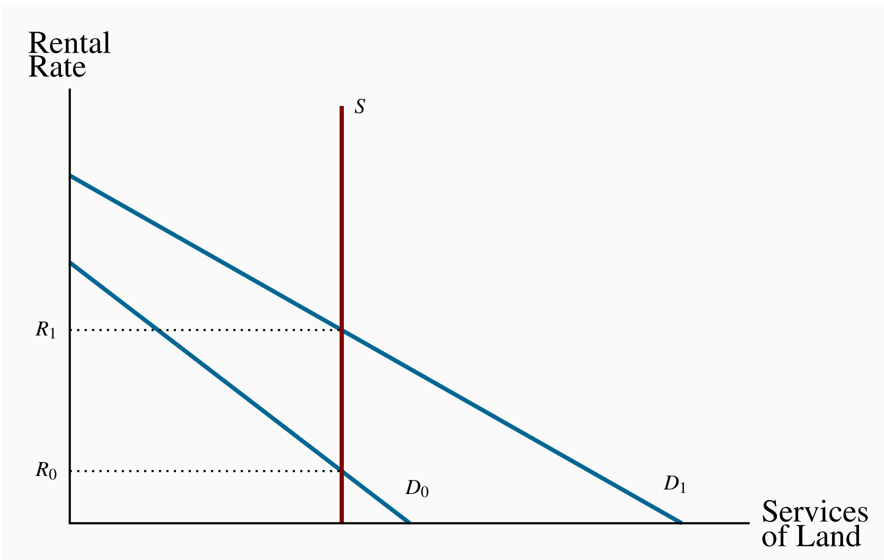
Note that the capital in this example is sector-specific. Drilling equipment cannot be easily redirected for use in other sectors. In contrast, earth movers can move from one sector of the economy to another with greater ease. An earth mover can be used to dig foundations for housing or commercial buildings; it can be used for strip mining; to build roads and bridges; to build tennis courts, golf courses and public parks. Such equipment may thus be moved to other sectors of the economy if in one particular sector the capital no longer can earn the required rental.

The prices of capital goods in the long run will be determined by the supply and demand for the services they provide. If the value of the services, as determined by supply and demand is high, then the price of assets will reflect this.

12.6 Land

Land is an input used in production, though is not a capital good in the way we defined capital goods earlier – production inputs that are themselves produced in the economy. Land is relatively fixed in supply *to the economy*, even in the long run. While this may not be literally true – the Netherlands reclaimed from the sea a great quantity of low-lying farmland, and fertilizers can turn marginal land into fertile land – it is a good approximation to reality. Figure 12.6 shows the derived demand D_0 for land services. With a fixed supply S, the equilibrium rental is R_0.

Figure 12.6: The market for land services



The supply of land is relatively fixed, and therefore the return to land is primarily demand determined. Shifts in demand give rise to differences in returns.

In contrast to this economy-wide perspective, consider now a retailer who rents space in a commercial mall. The area around the mall experiences a surge in development and more people are shopping and doing business there. The retailer finds that she sells more, but also finds that her rent increases on account of the additional demand for space by commercial enterprises in the area. Her landlord is able to charge a higher rent because so many potential clients wish to rent space in the area. Consequently, despite the additional commerce in the area, the retailer's profit increase will be moderated by the higher rents she must pay: The demand for retail space is a *derived demand*. The situation can be explained with reference to Figure 12.6 again. On account of growth in this area, the demand for retail space shifts from D_0 to D_1. Space in the area is restricted, and thus the vertical supply curve describes the supply side well. So with little or no possibility of higher prices bringing forth additional supply, the additional demand makes for a steep price (rent) increase.

Land has many uses and the returns to land must reflect this. Land in downtown Vancouver is priced higher than land in rural Saskatchewan. Land cannot be moved from the latter to the former location however, and therefore the rent differences represent an equilibrium. In contrast, land in downtown Winnipeg that is used for a parking lot may not be able to compete with the use of that land for office development. Therefore, for it to remain as a parking lot, the rental must reflect its high opportunity cost. This explains why parking fees in big US cities such as Boston or New York may run to $40 per day. If the parking owners could not obtain this fee, they could profitably sell the land to a developer. Ultimately it is the *value in its most productive use* that determines the price of land.

Key Terms

Demand for labour: a derived demand, reflecting the demand for the output of final goods and services.

Value of the marginal product is the marginal product multiplied by the price of the good produced.

Marginal revenue product of labour is the additional revenue generated by hiring one more unit of labour where the marginal revenue declines.

Monopsonist is the sole buyer of a good or service and faces an upward-sloping supply curve.

Participation rate: the fraction of the population in the working age group that joins the labour force.

The **labour force** is that part of the population either employed or seeking employment.

Unemployment rate: the fraction of the labour force actively seeking employment that is not employed.

Transfer earnings are the amount that an individual can earn in the next highest paying alternative job.

Rent is the excess remuneration an individual currently receives above the next best alternative. This alternative is the reservation wage.

Physical capital is the stock of produced goods that are inputs to the production of other goods and services.

Gross investment is the production of new capital goods and the improvement of existing capital goods.

Net investment is gross investment minus depreciation of the existing capital stock.

Depreciation is the annual change in the value of a physical asset.

Stock is the quantity of an asset at a point in time.

Flow is the stream of services an asset provides during a period of time.

Capital services are the production inputs generated by capital assets.

Rental rate: the cost of using capital services.

Asset price: the financial sum for which the asset can be purchased.

Present value of a stream of future earnings: the sum of each year's earnings divided by one plus the interest rate raised to the appropriate power.

Marginal product of capital is the output produced by one additional unit of capital services, with all other inputs being held constant.

Value of the marginal product of capital is the marginal product of capital multiplied by the price of the output it produces.

Required rental covers the sum of maintenance, depreciation and interest costs.

EXERCISES FOR CHAPTER 12

Exercise 12.1 Aerodynamics is a company specializing in the production of bicycle shirts. It has a fixed capital stock, and sells its shirts for \$20 each. It pays a weekly wage of \$400 per worker. Aerodynamics must maximize its profits by determining the optimal number of employees to hire. The marginal product of each worker can be inferred from the table below. Determine the optimal number of employees. [*Hint*: You must determine the VMP_L schedule, having first computed the MP_L.]

Employment	0	1	2	3	4	5	6
Total output	0	20	50	75	95	110	120
$\mathbf{MP_L}$							
$\mathbf{VMP_L}$							

Exercise 12.2 Suppose that, in Exercise 12.1 above, wages are not fixed. Instead the firm must pay \$50 more to employ each individual worker: The first worker is willing to work for \$250, the second for \$300, the third for \$350, etc. But once employed, each worker actually earns the same wage. Determine the optimal number of workers to be employed. [*Hint*: You must recognize that each worker earns the same wage; so when one additional worker is hired, the wage must increase to all workers employed.]

Exercise 12.3 Consider the following supply and demand equations for berry pickers. Demand: $W = 22 - 0.4L$; supply: $W = 10 + 0.2L$.

(a) For values of $L = 1, 5, 10, 15, \ldots, 30$, calculate the corresponding wage in each of the supply and demand functions.

(b) Using the data from part (a), plot and identify the equilibrium wage and quantity of labour.

(c) Illustrate in the diagram the areas defining transfer earnings and rent.

(d) Compute the transfer earnings and rent components of the total wage bill.

Exercise 12.4 The rows of the following table describe the income stream for three different capital investments. The income flows accrue in years 1 and 2. Only year 2 returns need to be discounted. The rate of interest is the first entry in each row, and the project cost is the final entry.

Interest rate	**Year 1**	**Year 2**	**Cost**
8%	8,000	9,000	16,000
6%	0	1,000	900
10%	4,000	5,000	11,000

(a) For each investment calculate the present value of the stream of services.

(b) Decide whether or not the investment should be undertaken.

Exercise 12.5 Nihilist Nicotine is a small tobacco farm in south-western Ontario. It has three plots of land, each with a different productivity, in that the annual yield differs across plots. The output from each plot is given in the table below. Each plot is the same size and requires 3 workers and one machine to harvest the leaves. The cost of these inputs is \$10,000. If the price of each kilogram of leaves is \$4, how many plots should be planted?

Land plot	Leaf yield in kilograms
One	3,000
Two	2,500
Three	2,000

Exercise 12.6 The timing of wine sales is a frequent problem encountered by vintners. This is because many red wines improve with age. Let us suppose you own a particular vintage and you envisage that each bottle should increase in value by 10% the first year, 9% the second year, 8% the third year, etc.

(a) Suppose the interest rate is 5%, for how many years would you hold the wine if there is no storage cost?

(b) If in addition to interest rate costs, there is a cost of storing the wine that equals 2% of the wine's value each year, for how many years would you hold the wine before selling?

Exercise 12.7 *Optional*: The industry demand for plumbers is given by the equation $W = 50 - 0.08L$, and there is a fixed supply of 300 qualified plumbers.

(a) Draw a diagram illustrating the supply, demand and equilibrium, knowing that the quantity intercept for the demand equation is 625.

(b) Solve the supply and demand equations for the equilibrium wage, W.

(c) If the plumbers now form a union, and supply their labour at a wage of \$30 per hour, illustrate the new equilibrium on your diagram and calculate the new level of employment.

Chapter 13

Human capital and the income distribution

In this chapter we will explore:

Individuals with different characteristics earn different amounts because their productivity levels differ. While it is convenient to work with a single marginal productivity of labour function to illustrate the functioning of the labour market, as we did in Chapter 12, for the most part wages and earnings vary by education level and experience, and sometimes by ethnicity and gender. In this chapter we develop an understanding of the sources of these differentials, and how they are reflected in the distribution of income.

13.1 Human capital

Human capital, HK, is the stock of knowledge and ability accumulated by a worker that determines future productivity and earnings. It depends on many different attributes – education, experience, intelligence, interpersonal skills etc. While human capital influences own earnings, it also impacts the productivity of the economy at large, and is therefore a vital force in determining long-run growth. Canada has been investing heavily in human capital in recent decades, and this suggests that future productivity and earnings will benefit accordingly.

> **Human capital** is the stock of knowledge and ability accumulated by a worker that determines future productivity and earning.

Several features of Canada's recent human capital accumulation are noteworthy. First, Canada's enrollment rate in post-secondary education now exceeds the US rate, and that of virtually every economy in the world. Second is the fact that the number of women in third-level institutions exceeds the number of men. Almost 60% of university students are women. Third, international testing of high-school students sees Canadian students performing well, which indicates that the quality of the Canadian educational system appears to be high. These are positive aspects of a system that frequently comes under criticism. Nonetheless, the distribution of income that emerges from market forces in Canada has become more unequal.

Let us now try to understand individuals' *economic* motivation for embarking on the accumulation of human capital, and in the process see why different groups earn different amounts. We start by analyzing the role of education and then turn to on-the-job training. At the outset, we recognize that many individuals acquire knowledge for its own sake or in order to improve the quality of their lives. Education can improve one's appreciation of art, literature and the sciences.

13.2 Productivity and education

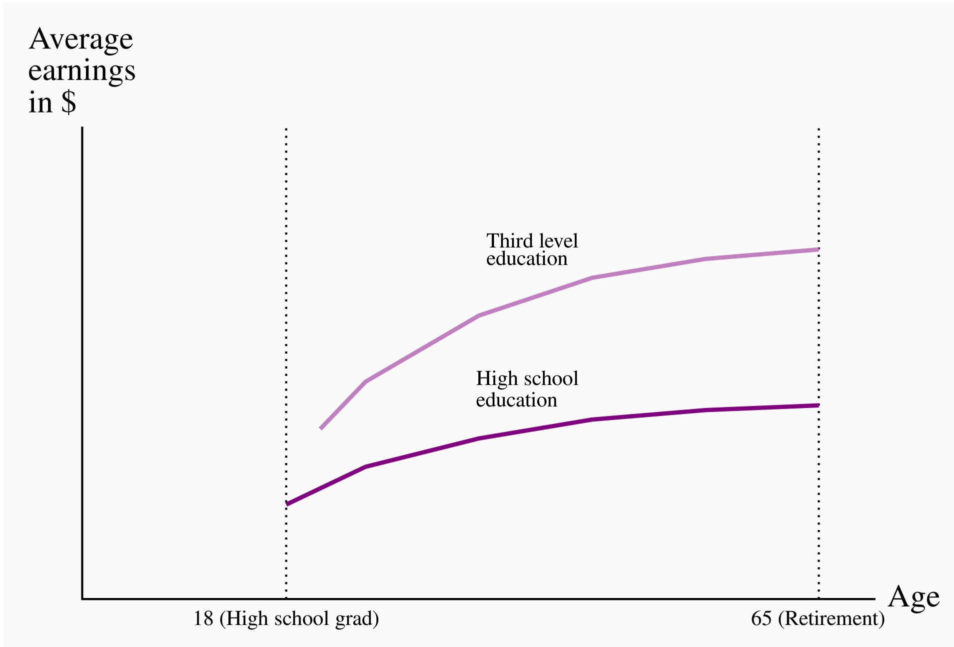
Human capital is the result of past investment that raises future incomes. A critical choice for individuals is to decide upon exactly how much additional human capital to accumulate. The cost of investing in another year of school is the *direct cost*, such as school fees, plus the *indirect*, or *opportunity*, *cost*, which can be measured by the foregone earnings during that extra year. The benefit of the additional investment is that the future flow of earnings is augmented. Consequently, wage differentials should reflect different degrees of education-dependent productivity.

Age-earnings profiles

Figure 13.1 illustrates two typical **age-earnings profiles** for individuals with different levels of education. These profiles define the typical pattern of earnings over time, and are usually derived by examining averages across individuals in surveys. Two aspects are clear: People with more education not only earn more, but the spread tends to grow with time. Less educated, healthy young individuals who work hard may earn a living wage but, unlike their more educated counterparts, they cannot look forward to a wage that rises substantially over time. More highly-educated individuals go into jobs and occupations that take a longer time to master: Lawyers, doctors and most professionals not only undertake more schooling than truck drivers, they also spend many years learning on the job, building up a clientele and accumulating expertise.

> **Age-earnings profiles** define the pattern of earnings over time for individuals with different characteristics.

Figure 13.1: Age-Earnings profiles by education level



Individuals with a higher level of education earn more than individuals with a 'standard' level of education. In addition, the differential grows over time.

The education premium

Individuals with different education levels earn different wages. The **education premium** is the difference in earnings between the more and less highly educated. Quantitatively, Professors Kelly Foley and David Green have recently proposed that the completion of a college or trade certification adds about 15% to one's income, relative to an individual who has completed high school. A Bachelor's degree brings a premium of 20-25%, and a graduate degree several percentage points more[1]. The failure to complete high school penalizes individuals to the extent of about 10%. These are average numbers, and they vary depending upon the province of residence, time period and gender. Nonetheless the findings underline that more human capital is associated with higher earnings. The earnings premium depends upon both the supply and demand of high HK individuals. *Ceteris paribus*, if high-skill workers are heavily in demand by employers, then the premium should be greater than if lower-skill workers are more in demand.

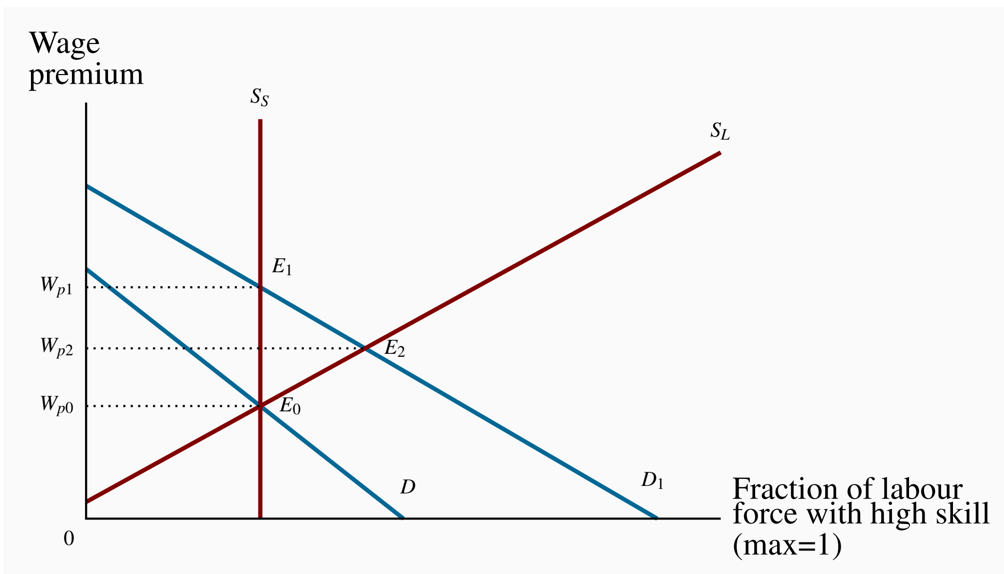
> **Education premium**: the difference in earnings between the more and less highly educated.

The distribution of earnings has become more unequal in Canada and the US in recent decades, and one reason that has been proposed for this development is that the modern economy demands

[1] Foley, K. and D. Green, 2015, "Why more education will not solve rising inequality (and may make it worse)" , *Institute for Research in Public Policy*, Montreal, Canada.

more high-skill workers; in particular that technological change has a bigger impact on productivity when combined with high-skill workers than with low-skill workers. Consider Figure 13.2 which contains supply and demand functions with a twist. We imagine that there are two types of labour: One with a high level of human capital, the other with a lower level. The vertical axis measures the wage *premium* of the high-education group (which can be measured in dollars or percentage terms), and the horizontal axis measures the *fraction of the total labour force that is of the high-skill type*. *D* is the *relative* demand for the high skill workers, in this example for the economy as a whole. There is some degree of substitution between high and low-skill workers in the modern economy. We do not propose that several low-skill workers can perform the work of one neuro-surgeon; but several individual households (low-skill) could complete their income tax submissions in the same time as one skilled tax specialist. In this example there is a degree of substitutability. In a production environment, a high-skill manager, equipped with technology and capital, can perform the tasks of several line workers.

Figure 13.2: The education/skill premium



A shift in demand increases the wage premium in the short run (from E_0 to E_1) by more than in the long run (to E_2). In the short run, the percentage of the labour force (S_S) that is highly skilled is fixed. In the long run it (S_L) is variable and responds to the wage premium.

The demand curve *D* defines the premium that employers (demanders) are willing to pay to the higher skill group. The negative slope indicates that if demanders were to employ a high proportion of skilled workers, the premium they would be willing to pay would be less than if they demanded a smaller share of high-skilled workers, and a larger share of lower-skilled workers. The wage premium for high HK individuals at any given time is determined by the intersection of supply and demand.

In the short run the make-up of the labour force is fixed, and this is reflected in the vertical supply curve S_s. The equilibrium is at E_0, and W_{p0} is the premium, or excess, paid to the higher-skill worker over the lower-skill worker. In the long run it is possible for the economy to change the composition of its labour supply: If the wage premium increases, more individuals will find it profitable to train as high-skill workers. That is to say, the fraction of the total that is high-skill increases. It follows that the long-run supply curve slopes upwards.

So what happens when there is an increase in the demand for high-skill workers relative to low-skill workers? The demand curve shifts upward to D_1, and the new equilibrium is at E_1. The supply mix is fixed in the short run, so there is an increase in the wage premium. But over time, some individuals who might have been just indifferent between educating themselves more and going into the workplace with lower skill levels now find it worthwhile to pursue further education. Their higher anticipated returns to the additional human capital they invest in now exceed the additional costs of more schooling, whereas before the premium increase these additional costs and benefits were in balance. In Figure 13.2 the new short-run equilibrium at E_1 has a corresponding wage premium of W_{p1}. In the long run, after additional supply has reached the market, the increased premium is moderated to W_{p2} at the equilibrium E_2.

This figure displays what many economists believe has happened in North America in recent decades: The demand for high HK individuals has increased, and the additional supply has not been as great. Consequently the wage premium for the high-skill workers has increased. As we describe later in this chapter, that is not the only perspective on what has happened.

Are students credit-constrained or culture-constrained?

The foregoing analysis assumes that students and potential students make rational decisions on the costs and benefits of further education and act accordingly. It also assumes implicitly that individuals can borrow the funds necessary to build their human capital: If the additional returns to further education are worthwhile, individuals should borrow the money to make the investment, just as entrepreneurs do with physical capital.

However, there is a key difference in the credit markets. If an entrepreneur fails in her business venture the lender will have a claim on the physical capital. But a bank cannot repossess a human being who drops out of school without having accumulated the intended human capital. Accordingly, the traditional lending institutions are frequently reluctant to lend the amount that students might like to borrow—students are credit constrained. The sons and daughters of affluent families therefore find it easier to attend university, because they are more likely to have a supply of funds domestically. Governments customarily step into the breach and supply loans and bursaries to students who have limited resources. While funding frequently presents an obstacle to attending a third-level institution, a stronger determinant of attendance is the education of the parents, as detailed in Application Box 13.1.

Application Box 13.1: Parental education and university attendance in Canada

The biggest single determinant of university attendance in the modern era is parental education. A recent study* of who goes to university examined the level of parental education of young people 'in transition' – at the end of their high school – for the years 1991 and 2000.

For the year 2000 they found that, if a parent had not completed high school, there was just a 12% chance that their son would attend university and an 18% chance that a daughter would attend. In contrast, for parents who themselves had completed a university degree, the probability that a son would also attend university was 53% and for a daughter 62%. Hence, the probability of a child attending university was roughly four times higher if the parent came from the top educational category rather than the bottom category! Furthermore the authors found that this probability gap opened wider between 1991 and 2000.

In the United States, Professor Sear Reardon of Stanford University has followed the performance of children from low-income households and compared their achievement with children from high-income households. He has found that the achievement gap between these groups of children has increased substantially over the last three decades. The reason for this growing separation is not because children from low-income households are performing worse in school, it is because high-income parents invest much more of their time and resources in educating their children, both formally in the school environment, and also in extra-school activities.

*Finnie, R., C. Laporte and E. Lascelles. "Family Background and Access to Post-Secondary Education: What Happened in the Nineties?" Statistics Canada Research Paper, Catalogue number 11F0019MIE-226, 2004

Reardon, Sean, "The Great Divide", New York Times, April 8, 2015.

13.3 On-the-job training

As is clear from Figure 13.1, earnings are raised both by education and experience. Learning on the job is central to the age-earnings profiles of the better educated, and is less important for those with lower levels of education. **On-the-job training** improves human capital through work experience. If on-the-job training increases worker productivity, who should pay for this learning – firms or workers? To understand who should pay, we distinguish between two kinds of skills: **Firm-specific skills** that raise a worker's productivity in a particular firm, and **general skills** that enhance productivity in many jobs or firms.

Firm-specific HK could involve knowing how particular components of a somewhat unique production structure functions, whereas general human capital might involve an understanding of engineering or architectural principles that can be applied universally. As for who should pay for the accumulation of skills: An employer should be willing to undertake most of the cost of firm-specific skills, because they are of less value to the worker should she go elsewhere. Firms offering general or transferable training try to pass the cost on to the workers, perhaps by offering a wage-earnings profile that starts very low, but that rises over time. Low-wage apprenticeships are examples. Hence, whether an employee is a medical doctor in residence, a plumber in an apprenticeship or a young lawyer in a law partnership, she 'pays' for the accumulation of her portable HK by facing a low wage when young. Workers are willing to accept such an earnings profile because their projected future earnings will compensate for lower initial earning.

> **On-the-job training** improves human capital through work experience.
>
> **Firm-specific skills** raise a worker's productivity in a particular firm.
>
> **General skills** enhance productivity in many jobs or firms.

13.4 Education as signalling

An alternative view of education springs from the theory of signalling. This is a provocative theory that proposes education may be worthwhile, even if it generates little additional skills. The theory recognizes that individuals possess different abilities. However, firms cannot easily recognize the more productive workers without actually hiring them and finding out *ex-post* – sometimes a costly process. **Signalling theory** says that, in pursuing more education, people who know they are more capable thereby send a signal to potential employers that they are the more capable workers. Education therefore **screens** out the low-productivity workers from the high-productivity (more educated) workers. Firms pay more to the latter, because firms know that the high-ability workers are those with the additional education.

> **Signalling** is the decision to undertake an action in order to reveal information.
>
> **Screening** is the process of obtaining information by observing differences in behaviour.

To be effective, the process must separate the two types. Why don't lower-ability workers go to university and pretend they are of the high-ability type? Primarily because that strategy could backfire: Such individuals are less likely to succeed at school and there are costs associated with school in the form of school fees, books and foregone earnings. While they may have lower innate skills, they are likely smart enough to recognize a bad bet.

On balance economists believe that further education does indeed add to productivity, although there may be an element of screening present: An engineering degree (we should hope) increases an individual's understanding of mechanical forces so that she can design a bridge that will not collapse, in addition to telling a potential employer that the student is smart!

Finally, it should be evident that if education raises productivity, it is also good for society and the economy at large.

13.5 Education returns and quality

How can we be sure that further education really does generate the returns, in the form of higher future incomes, to justify the investment? For many years econometricians proposed that an extra year of schooling might offer a return in the region of 10% – quite a favourable return in comparison with what is frequently earned on physical capital. Doubters then asked if the econometric estimation might be subject to bias – what if the additional earnings of those with more education are simply attributable to the fact that it is the innately more capable individuals who both earn more and who have more schooling? And since we cannot observe who has more innate ability, how can we be sure that it is the education itself, rather than just differences in ability, that generate the extra income?

This is a classical problem in inference: Does correlation imply causation? The short answer to this question is that education economists are convinced that the time invested in additional schooling does indeed produce additional rewards, even if it is equally true that individuals who are innately smarter do choose to invest in that way. Furthermore, it appears that the returns to graduate education are higher than the returns to undergraduate education.

What can be said of the quality of different educational systems? Are educational institutions in different countries equally good at producing knowledgeable students? Or, viewed another way: Has a grade nine student in Canada the same skill set as a grade nine student in France or Hong Kong? An answer to this question is presented in Table 13.1, which contains results from the Program for International Student Assessment (PISA) – an international survey of 15-year old student abilities in mathematics, science and literacy. This particular table presents the results for a sample of the countries that were surveyed. The results indicate that Canadian students perform well in all three dimensions of the test.

Table 13.1: Mean scores in PISA tests

Country	Math	Science	Reading
Australia	494	510	503
Austria	497	495	485
Belgium	507	502	499
Canada	516	528	527
Denmark	511	502	500
Finland	511	531	526
France	493	495	499
Germany	506	509	509
Greece	454	455	467
Hong Kong	548	523	527
Ireland	504	503	521
Italy	490	481	485
Japan	532	538	516
Korea	524	516	517
Mexico	408	416	423
New Zealand	495	513	509
Norway	502	498	513
Spain	486	493	496
Sweden	494	493	500
Switzerland	521	506	492
Turkey	420	425	428
United States	470	496	497
United Kingdom	492	509	498

Source: https://www.oecd.org/pisa/pisa-2015-results-in-focus.pdf

An interesting paradox arises at this point: If productivity growth in Canada has lagged behind some other economies in recent decades, as many economists believe, how can this be explained if Canada produces many well-educated high-skill workers? The answer may be that there is a considerable time lag before high participation rates in third-level education and high quality make themselves felt on the national stage in the form of elevated productivity. The evidence suggests a good productivity future, in so far as it depends upon human capital. At the same time, if investment in human capital is not matched by investment in physical capital then the human capital may not be able to perform to its ability.

13.6 Discrimination

Wage differences are a natural response to differences in human capital. But we frequently observe wage differences that might be discriminatory. For example, women on average earn less than men with similar qualifications; older workers may be paid less than those in their prime years; immigrants may be paid less than native-born Canadians, and ethnic minorities may be paid less than traditional white workers. The term **discrimination** describes an earnings differential that is attributable to a trait other than human capital.

If two individuals have the same HK, in the broadest sense of having the same capability to perform a particular task, then a wage premium paid to one represents discrimination. Correctly measured then, the discrimination premium between individuals from these various groups is the differential in earnings after correcting for HK differences. Thousands of studies have been undertaken on discrimination, and most conclude that discrimination abounds. Women, particularly those who have children, are paid less than men, and frequently face a 'glass ceiling' – a limit on their promotion possibilities within organizations.

> **Discrimination** implies an earnings differential that is attributable to a trait other than human capital.

In contrast, women no longer face discrimination in university and college admissions, and form a much higher percentage of the student population than men in many of the higher paying professions such as medicine and law. Immigrants to Canada also suffer from a wage deficit. This is especially true for the most recent cohorts of working migrants who now come predominantly, not from Europe, as was once the case, but from China, South Asia, Africa and the Caribbean. For similarly-measured HK as Canadian-born individuals, these migrants frequently have an initial wage deficit of 30%, and require a period of more than twenty years to catch-up. How much of this differential might be due to the quality of the education or human capital received abroad is difficult to determine.

13.7 The income distribution

How does all of our preceding discussion play out when it comes to the income distribution? That is, when we examine the incomes of all individuals or households in the economy, how equally or unequally are they distributed?

The study of inequality is a critical part of economic analysis. It recognizes that income differences that are in some sense 'too large' are not good for society. Inordinately large differences can reflect poverty and foster social exclusion and crime. Economic growth that is concentrated in the hands of the few can increase social tensions, and these can have economic as well as social or psychological costs. Crime is one reflection of the divide between 'haves' and 'have-nots'. It is economically costly; but so too is child poverty. Impoverished children rarely achieve their social or economic potential and this is a loss both to the individual and society at large.

In this section we will first describe a subset of the basic statistical tools that economists use to measure inequality. Second, we will examine how income inequality has evolved in recent decades. We shall see that, while the picture is complex, market income inequality has indeed increased in Canada. Third, we shall investigate some of the proposed reasons for the observed increase in inequality. Finally we will examine if the government offsets the inequality that arises from the marketplace through its taxation and redistribution policies.

It is to be emphasized that income inequality is just one proximate measure of the distribution of wellbeing. The extent of poverty is another such measure. Income is not synonymous with happiness but, that being said, income inequality can be computed reliably, and it provides a good measure of households' control over economic resources.

Theory and measurement

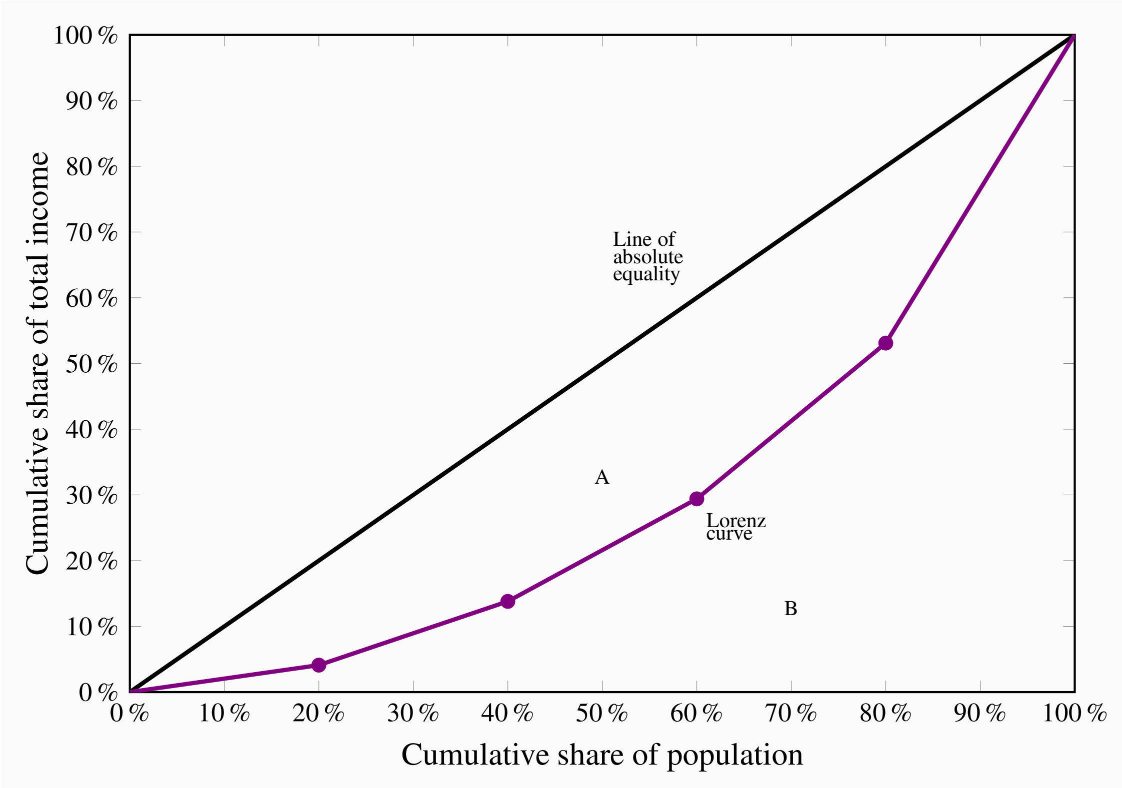
Let us rank the market incomes of all households in the economy from poor to rich, and categorize this ordering into different quantiles or groups. With five such quantiles the shares are called *quintiles*. The richest group forms the highest quintile, while the poorest group forms the lowest quintile. Such a representation is given in Table 13.2. The first numerical column displays the income in each quintile as a percentage of total income. If we wanted a finer breakdown, we could opt for decile (ten), or even vintile (twenty) shares, rather than quintile shares. These data can be graphed in a variety of ways. Since the data are in share, or percentage, form, we can compare, in a meaningful manner, distributions from economies that have different average income levels.

Table 13.2: Quintile shares of total family income in Canada, 2011

	Quintile share of total income	Cumulative share
First quintile	4.1	4.1
Second quintile	9.6	13.7
Third quintile	15.3	29.0
Fourth quintile	23.8	52.8
Fifth quintile	47.2	100.0
Total	100	

Source: Statistics Canada, CANSIM Matrix 2020405. These combinations are represented by the circles in the figure.

Figure 13.3: Gini index and Lorenz curve



The more equal are the income shares, the closer is the Lorenz curve to the diagonal line of equality. The Gini index is the ratio of the area A to the area (A+B). The Lorenz curve plots the cumulative percentage of total income against the cumulative percentage of the population.

An informative way of presenting these data graphically is to plot the cumulative share of income against the cumulative share of the population. This is given in the final column, and also presented graphically in Figure 13.3. The bottom quintile has 4.1% of total income. The bottom two quintiles together have $13.7\%(4.1\% + 9.6\%)$, and so forth. By joining the coordinate pairs represented by the circles, a **Lorenz curve** is obtained. Relative to the diagonal line it is a measure of how unequally incomes are distributed: If everyone had the same income, each 20% of the population would have 20% of total income and by joining the points for such a distribution we would get a straight diagonal line joining the corners of the box. In consequence, if the Lorenz curve is further from the line of equality the distribution is less equal than if the Lorenz curve is close to the line of equality.

> **Lorenz curve** describes the cumulative percentage of the income distribution going to different quantiles of the population.

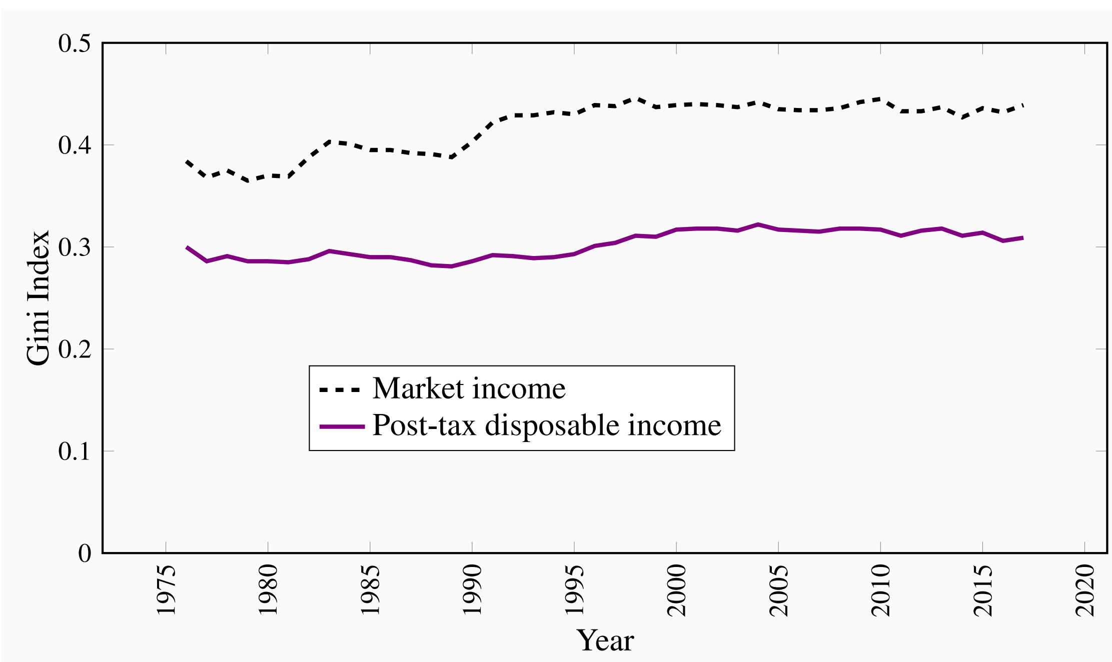
This suggests that the area A relative to the area (A + B) forms a measure of inequality in the income distribution. This fraction obviously lies between zero and one, and it is called the **Gini index**. A larger value of the Gini index indicates that inequality is greater. We will not delve into the mathematical formula underlying the Gini, but for this set of numbers its value is 0.4.

> **Gini index**: a measure of how far the Lorenz curve lies from the line of equality. Its maximum value is one; its minimum value is zero.

The Gini index is what is termed *summary index* of inequality – it encompasses a lot of information in one number. There exist very many other such summary statistics.

It is important to recognize that very different Gini index values emerge for a given economy by using different income definitions of the variable going into the calculations. For example, the quintile shares of the *earnings of individuals* rather than the *incomes of households* could be very different. Similarly, the shares of income *post tax* and *post transfers* will differ from their shares on a *pre-tax*, *pre-transfer* basis.

Figure 13.4: Gini index Canada 1976–2015



Source: Statistics Canada, CANSIM Table 206-0033

Figure 13.4 contains Gini index values for two different definitions of income from 1976 to 2011. The upper line represents the Gini index values for households where the income measure is market income; the lower line defines the Gini values when income is defined as post-tax and post-transfer incomes. The latter income measure deducts taxes paid and adds income such as Employment Insurance or Social Assistance benefits. Two messages emerge from this graphic: The first is that the distribution of market incomes displays more inequality than the distribution of incomes after the government has intervened. In the latter case incomes are more equally distributed than in the former. The second message to emerge is that inequality increased over time – the Gini values are larger in the years after approximately 2000 than in the earlier years, although the increase in market income inequality is greater than the increase in income inequality based on a 'post-government' measure of income.

This is a very brief description of recent events. It is also possible to analyze inequality among women and men, for example, as well as among individuals and households. But the essential message remains clear: Definitions are important; in particular the distinction between incomes generated in the market place and incomes after the government has intervened through its tax and transfer policies.

Application Box 13.2: The very rich

McMaster University Professor Michael Veall and his colleague Emmanuel Saez, from University of California, Berkeley, have examined the evolution of the top end of the Canadian earnings distribution in the twentieth century. Using individual earnings from a database built upon tax returns, they show how the share of the very top of the distribution declined in the nineteen thirties and forties, remained fairly stable in the decades following World War II, and then increased from the eighties to the present time. The increase in share is particularly strong for the top 1% and even stronger for the top one tenth of the top 1%. These changes are driven primarily by changes in earnings, not on stock options awarded to high-level corporate employees. The authors conclude that the change in this region of the distribution is attributable to changes in social norms. Whereas, in the nineteen eighties, it was expected that a top executive would earn perhaps a half million dollars, the 'norm' has become several million dollars in the present day. Such high remuneration became a focal point of public discussion after so many banks in the United States in 2008 and 2009 required government loans and support in order to avoid collapse. It also motivated the many 'occupy' movements of 2011 and 2012, and the US presidential race in 2019.

Saez, E. and M. Veall. "The evolution of high incomes in Canada, 1920-2000." Department of Economics research paper, McMaster University, March 2003.

In the international context, Canada is neither a strongly egalitarian economy nor one characterized by great income inequality. OECD data indicate that the economies with the lowest Gini index values are the Czech and Slovak republics and Iceland, with values in the neighborhood of 0.25 based on a post-government measure of income. Canada has a Gini index of .31, the US a value of .38 and at the upper end are economies such as Mexico and Chile with values of .47 (`https://data.oecd.org/inequality/income-inequality.htm`).

Economic forces

The increase in inequality of earnings in the market place in Canada has been reflected in many other developed economies – to a greater degree in the US and to a lesser extent in some European economies. Economists have devoted much energy to studying why, and as a result there are several accepted reasons.

Younger workers and those with lower skill levels have faired poorly in the last three decades. **Globalization and out-sourcing** have put pressure on low-end wages. In effect the workers in the lower tail of the distribution are increasingly competing with workers from low-wage less-developed economies. While this is a plausible causation, the critics of the perspective point out that wages at the bottom have fallen not only for those workers who compete with overseas workers in manufacturing, but also in the domestic services sector right across the economy. Obviously the workers at *McDonalds* have not the same competition from low-wage economies as workers who assemble toys.

A competing perspective is that it is **technological change** that has enabled some workers to do better than others. In explaining why high wage workers in many economies have seen their wages increase, whereas low-wage workers have seen a relative decline, the technological change hypothesis proposes that the form of recent technological change is critical: Change has been such as to require other complementary skills and education in order to benefit from it. For example, the introduction of computer-aided design technology is a benefit to workers who are already skilled and earning a high wage: *Existing high skills and technological change are complementary*. Such technological change is therefore different from the type underlying the production line. Automation in the early twentieth century in Henry Ford's plants improved the wages of lower skilled workers. But in the modern economy it is the highly skilled rather than the low skilled that benefit most from innovation.

A third perspective is that key **institutional changes** manifested themselves in the eighties and nineties, and these had independent impacts on the distribution. In particular, declines in the extent of unionization and changes in the minimum wage had significant impacts on earnings in the middle and bottom of the distribution: If unionization declines or the minimum wage fails to keep up with inflation, these workers will suffer. An alternative 'institutional' player is the government: In Canada the federal government became slightly less supportive, or 'generous', with its array of programs that form Canada's social safety net in the nineteen nineties. This tightening goes some way to explaining the modest inequality increase in the post-government income distribution in Figure 13.3 at this time. Nonetheless, most Canadian provincial governments increased the legal minimum wage in the first decade of the new millennium by substantially more than the rate of inflation. This meant that the economy's low-income workers did not fall further behind.

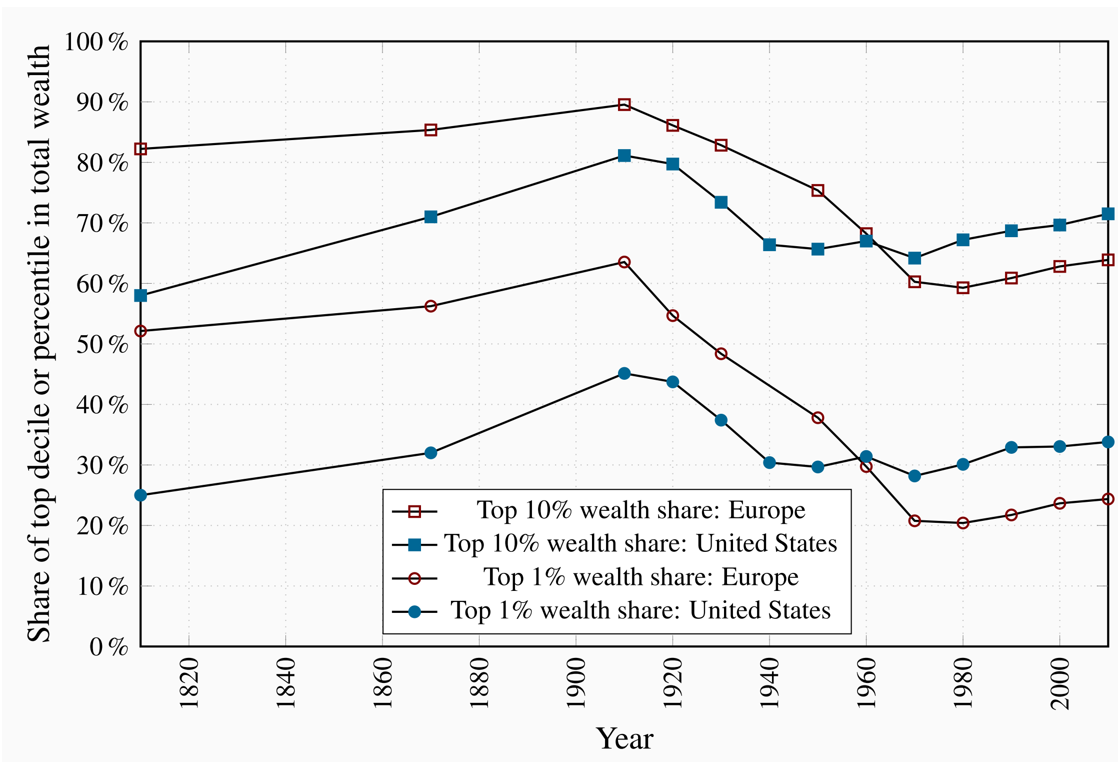
We conclude this overview of distributional issues by pointing out that we have not analyzed the distribution of wealth. Wealth too represents purchasing power, and it is wealth rather than income flows that primarily distinguishes Warren Buffet, Mark Zuckerberg and Bill Gates from the rest of us mortals. A detailed treatment of wealth inequality is beyond the scope of this book. We describe briefly, in the final section, the recent contribution of Thomas Piketty to the inequality debate.

13.8 Wealth and capitalism

In an insightful and popular study of capital accumulation, from both a historical and contemporary perspective, Thomas Piketty draws our attention to the enormous inequality in the distribution of wealth and explores what the future may hold in his book *Capital in the Twenty-First Century*.

The distribution of wealth is universally more unequal than the distribution of incomes or earnings. Gini coefficients in the neighbourhood of 0.8 are commonplace in developed economies. In terms of shares of the wealth pie, such a magnitude may imply that the top 1% of wealth holders own more than one third of all of an economy's wealth, that the top decile may own two-thirds of all wealth, and that the remaining one third is held by the 'bottom 90%'. And within this bottom 90%, virtually all of the remaining wealth is held by the 40% of the population below the top decile, leaving only a few percent of all wealth to the bottom 50% of the population.

Figure 13.5: Wealth inequality in Europe and the US, 1810-2010



Source: `http://piketty.pse.ens.fr/en/capital21c2`

While such an unequal holding pattern may appear shockingly unjust, Piketty informs us that current wealth inequality is not as great in most economies as it was about 1900. Figure 13.5 (Piketty 10.6) above is borrowed from his web site. Wealth was more unequally distributed in Old World Europe than New World America a century ago, but this relativity has since been reversed. A great transformation in the wealth holding pattern of societies in the twentieth century took the form of the emergence of a 'patrimonial middle class', by which he means the emergence of substantial wealth holdings on the part of that 40% of the population below the top decile. This development is noteworthy, but Piketty warns that the top percentiles of the wealth distribution may be on their way to controlling a share of all wealth close to their share in the early years of the twentieth century. He illustrates that two elements are critical to this prediction; first is the rate of growth in the economy relative to the return on capital, the second is inheritances.

To illustrate: Imagine an economy with very low growth, and where the owners of capital obtain an annual return of say 5%. If the owners merely maintain their capital intact and consume the remainder of this 5%, then the pattern of wealth holding will continue to be stable. However, if the holders of wealth can reinvest from their return an amount more than is necessary to replace depreciation then their wealth will grow. And if labour income in this economy is almost static on account of low overall growth, then wealth holders will secure a larger share of the economic pie. In contrast, if economic growth is significant, then labour income may grow in line with income from capital and inequality may remain stable. This summarizes Piketty's famous $(r-g)$ law – inequality depends upon the difference between the return on wealth and the growth rate of the economy. This potential for an ever-expanding degree of inequality is magnified when the stock of capital in the economy is large.

Consider now the role of inheritances. That is to say, do individuals leave large or small inheritances when they die, and how concentrated are such inheritances? If individual wealth accumulation patterns are generated by a desire to save for retirement and old age – during which time individuals decumulate by spending their assets – such motivation should result in small bequests being left to following generations. In contrast, if individuals who are in a position to do so save and accumulate, not just for their old age, but because they have dynastic preferences, or if they take pleasure simply from the ownership of wealth, or even if they are very cautious about running down their wealth in their old age, then we should see substantial inheritances passed on to the sons and daughters of these individuals, thereby perpetuating, and perhaps exacerbating, the inequality of wealth holding in the economy.

Piketty shows that in fact individuals who save substantial amounts tend to leave large bequests; that is they do not save purely for life-cycle motives. In modern economies the annual amount of bequests and gifts from parents to children falls in the range of 10% to 15% of annual GDP. This may grow in future decades, and since wealth is highly concentrated, these bequests in turn are concentrated among a small number of the following generation – inequality is transmitted from one generation to the next.

As a final observation, if we consider the distribution of income and wealth together, particularly at the very top end, we can see readily that a growing concentration of *income* among the top 1% should ultimately translate itself into greater *wealth* inequality. This is because top earners can save more easily than lower earners. To compound matters, if individuals who inherit wealth also tend to inherit more human capital from their parents than others, the concentration of income and wealth may become yet stronger.

The study of distributional issues in economics has probably received too little attention in the modern era. Yet it is vitally important both in terms of the well-being of the individuals who constitute an economy and in terms of adherence to social norms. Given that utility declines with additions to income and wealth, transfers from those at the top to those at the bottom have the potential to increase total utility in the economy. Furthermore, an economy in which justice is seen to prevail—in the form of avoiding excessive inequality—is more likely to achieve a higher degree of social coherence than one where inequality is large.

Key Terms

Human capital is the stock of expertise accumulated by a worker that determines future productivity and earnings.

Age-earnings profiles define the pattern of earnings over time for individuals with different characteristics.

Education premium: the difference in earnings between the more and less highly educated.

On-the-job training improves human capital through work experience.

Firm-specific skills raise a worker's productivity in a particular firm.

General skills enhance productivity in many jobs or firms.

Signalling is the decision to undertake an action in order to reveal information.

Screening is the process of obtaining information by observing differences in behaviour.

Discrimination implies an earnings differential that is attributable to a trait other than human capital.

Lorenz curve describes the cumulative percentage of the income distribution going to different quantiles of the population.

Gini index: a measure of how far the Lorenz curve lies from the line of equality. Its maximum value is one; its minimum value is zero.

Exercises for Chapter 13

Exercise 13.1 Georgina is contemplating entering the job market after graduating from high school. Her future lifespan is divided into two phases: An initial one during which she may go to university, and a second when she will work. Since dollars today are worth more than dollars in the future she discounts the future by 20%, that is the value today of that future income is the income divided by 1.2. By going to university and then working she will earn (i) -$60,000; (ii) $600,000. The negative value implies that she will incur costs in educating herself in the first period. In contrast, if she decides to work for both periods she will earn $30,000 in the first period and $480,000 in the second.

(a) If her objective is to maximize her lifetime earnings, should she go to university or enter the job market immediately?

(b) If instead of discounting the future at the rate of 20%, she discounts it at the rate of 50%, what should she do?

Exercise 13.2 Imagine that you have the following data on the income distribution for two economies.

	Quintile share of total income	
First quintile	4.1	3.0
Second quintile	9.6	9.0
Third quintile	15.3	17.0
Fourth quintile	23.8	29.0
Fifth quintile	47.2	42.0
Total	100	100

(a) On graph paper, or in a spreadsheet program, plot the Lorenz curves corresponding to the two sets of quintile shares. You must first compute the cumulative shares as we did for Figure 13.3.

(b) Can you say, from a visual analysis, which distribution is more equal?

Exercise 13.3 The distribution of income in the economy is given in the table below. The first numerical column represents the dollars earned by each quintile. Since the numbers add to 100 you can equally think of the dollar values as shares of the total pie. In this economy the government changes the distribution by levying taxes and distributing benefits.

Quintile	Gross income $m	Taxes $m	Benefits $m
First	4	0	9
Second	11	1	6
Third	19	3	5
Fourth	26	7	3
Fifth	40	15	3
Total	100	26	26

(a) Plot the Lorenz curve for gross income to scale.

(b) Now subtract the taxes paid and add the benefits received by each quintile. Check that the total income is still $100. Calculate the cumulative income shares and plot the resulting Lorenz curve. Can you see that taxes and benefits reduce inequality?

Exercise 13.4 Consider two individuals, each facing a 45 year horizon at the age of 20. Ivan decides to work immediately and his earnings path takes the following form: Earnings $= 20,000 + 1,000t - 10t^2$, where the t is time, and it takes on values from 1 to 25, reflecting the working lifespan.

(a) In a spreadsheet enter values $1 \dots 25$ in the first column and then compute the value of earnings in each of the 25 years in the second column using the earnings equation.

(b) John decides to study some more and only earns a part-time salary in his first few years. He hopes that the additional earnings in future years will compensate for that. His function is given by $10,000 + 2,000t - 12t^2$. In the same spreadsheet compute his annual earnings for 25 years.

(c) Plot the two earnings functions you have computed using the 'charts' feature of Excel. Does your graph indicate that John passes Ivan between year 10 and year 11?

Exercise 13.5 In the short run one half of the labour force has high skills and one half low skills (in terms of Figure 13.2 this means that the short-run supply curve is vertical at 0.5). The relative demand for the high-skill workers is given by $W = 40 \times (1 - f)$, where W is the wage premium and f is the fraction that is skilled. The premium is measured in percent and f has a maximum value of 1. The W function thus has vertical and horizontal intercepts of $\{40, 1\}$.

(a) Illustrate the supply and demand curves graphically, and illustrate the skill premium going to the high-skill workers in the short run by determining the value of W when $f = 0.5$.

(b) If demand increases to $W = 60 \times (1 - f)$ what is the new premium? Illustrate your answer graphically.

Exercise 13.6 Consider the foregoing problem in a long-run context, when the fraction of the labour force that is high-skilled is more elastic with respect to the premium. Let this long-run relative supply function be $W = 40 \times f$.

(a) Graph this long-run supply function and verify that it goes through the same initial equilibrium as in Exercise 13.5.

(b) Illustrate the long run and short run on the same diagram.

(c) What is the numerical value of the premium in the long run after the increase in demand? Illustrate graphically.

Part Six

Government and Trade

Governments play a major role in virtually every economy. They account for one third or more of national product. In addition to providing a legal and constitutional framework and implementing the law, governments can moderate and influence the operation or markets, provide public goods and furnish missing information. Governments distribute and redistribute widely and supply major services such as health and education. Modern economies could not function without a substantial role for their governments. These roles are explored and developed in Chapter 14.

The final chapter looks outwards. Canada is an open economy, with a high percentage of its production imported and exported. We explore the theory of absolute and comparative advantage and illustrate the potential for consumer gains that trade brings. We analyze barriers to trade such as tariffs and quotas and end with an overview of the World's major trading groups and institutions.

Chapter 14

Government

In this chapter we will explore:

Governments have a profound impact on economies. The economies of Scandinavia are very different from those in North America. North and South Korea are night and day, even though they were identical several decades ago. Canada and Argentina were very similar in the early decades of the twentieth century. Both had abundant space, natural resources and migrants. Today Canada is one of the most prosperous economies in the world, while Argentina struggles as a middle-income economy.

Governments are not peripheral to the marketplace. They can facilitate or hinder the operation of markets. They can ignore poverty or implement policies to support those on low income and those who are incapacitated. Governments can treat the economy as their fiefdoms, as has been the case for decades in many underdeveloped economies. By assuming the role of warlords, local governors inhibit economic development, because the fruits of investment and labour are subject to capture by the ruling power elite.

In Canada we take for granted the existence of a generally benign government that serves the economy, rather than one which expects the economy to serve it. The separation of powers, the existence of a constitution, property rights, a police force, and a free press are all crucial ingredients in the mix that ferments economic development and a healthy society.

The analysis of government is worthy not just of a full course, but a full program of study. Accordingly, our objective in this chapter must be limited. We begin by describing the various ways in which markets may be inadequate and what government can do to remedy these deficiencies. Next we describe the size and scope of government in Canada, and define the sources of government revenues. On the expenditure side, we emphasize the redistributive and transfer roles that are played by Canadian governments. Tax revenues, particularly at the federal level, go predominantly to transfers rather than the provision of goods and services. Finally we examine how governments seek to control, limit and generally influence the marketplace: How do governments foster the operation of markets in Canada? How do they attempt to limit monopolies and cartels? How do they attempt to encourage the entry of new producers and generally promote a market structure that is conducive to competition, economic growth and consumer well-being?

14.1 Market failure

Markets are fine institutions when all of the conditions for their efficient operation are in place. In Chapter 5 we explored the meaning of efficient resource allocation, by developing the concepts of consumer and producer surpluses. But, while we have emphasized the benefits of efficient resource allocation in a market economy, there are many situations where markets deliver inefficient outcomes. Several problems beset the operation of markets. The principal sources of *market failure* are: *Externalities, public goods, asymmetric information,* and the *concentration of power*. In addition markets may produce outcomes that are *unfavourable* to certain groups – perhaps those on low incomes. The circumstances described here lead to what is termed **market failure**.

Market failure defines outcomes in which the allocation of resources is not efficient.

Externalities

A negative externality is one resulting, perhaps, from the polluting activity of a producer, or the emission of greenhouse gases into the atmosphere. A positive externality is one where the activity of one individual confers a benefit on others. An example here is where individuals choose to get immunized against a particular illness. As more people become immune, the lower is the probability that the illness can propagate itself through hosts, and therefore the greater the benefits to those not immunized.

Solutions to these market failures come in several forms: Government taxes and subsidies, or quota systems that place limits on the production of products generating externalities. Such solutions were explored in Chapter 5. Taxes on gasoline discourage its use and therefore reduce the emission of poisons into the atmosphere. Taxes on cigarettes and alcohol lower the consumption of goods that may place an additional demand on our publicly-funded health system. The provision of free, or low-cost, immunization against specific diseases to children benefits the whole population.

These measures attempt to *compensate for the absence of a market* in certain activities. Producers may not wish to pay for the right to emit pollutants, and consequently if the government steps in to counter such an externality, the government is effectively implementing a solution to the missing market.

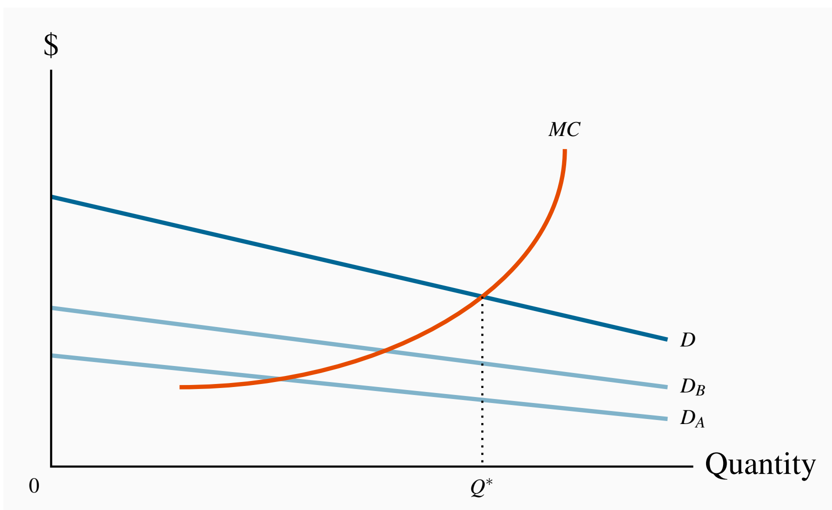
Public goods

Public goods are sometimes called collective consumption goods, on account of their non-rivalrous and non-excludability characteristics. For example, if the government meteorological office provides daily forecasts over the nation's airwaves, it is no more expensive to supply that information to one million than to one hundred individuals in the same region. Its provision to one is not rivalrous with its provision to others – in contrast to private goods that cannot be 'consumed' simultaneously by more than one individual. In addition, it may be difficult to exclude certain individuals from receiving the information.

> **Public goods** are non-rivalrous, in that they can be consumed simultaneously by more than one individual; additionally they may have a non-excludability characteristic.

Examples of such goods and services abound: Highways (up to their congestion point), street lighting, information on trans-fats and tobacco, or public defence provision. Such goods pose a problem for private markets: If it is difficult to exclude individuals from their consumption, then potential private suppliers will likely be deterred from supplying them because the suppliers cannot generate revenue from *free-riders*. Governments therefore normally supply such goods and services. But how much should governments supply? An answer is provided with the help of Figure 14.1.

Figure 14.1: Optimal provision of a public good



The total demand for the public good D is the vertical sum of the individual demands D_A and D_B. The optimal provision is where the MC equals the aggregate marginal valuation, as defined by the demand curve D. At the optimum Q^*, each individual is supplied the same amount of the public good.

This is a supply-demand diagram with a difference. The supply side is conventional, with the *MC* of production representing the supply curve. An efficient use of the economy's resources, we already know, dictates that an amount should be produced so that the cost at the margin equals the benefit to consumers at the margin. In contrast to the total market demand for private goods, which is obtained by summing individual demands horizontally, the demand for public goods is obtained by summing individual demands *vertically*.

Figure 14.1 depicts an economy with just two individuals whose demands for street lighting are given by D_A and D_B. These demands reveal the value each individual places on the various output levels of the public good, measured on the x-axis. However, since each individual can consume the public good *simultaneously*, the aggregate value of any output produced is the *sum of each individual valuation*. The valuation in the market of any quantity produced is therefore the vertical sum of the individual demands. D is the vertical sum of D_A and D_B, and the optimal output is Q^*. At this equilibrium each individual consumes the same quantity of street lighting, and the MC of the last unit supplied equals the value placed upon it by society – both individuals. Note that this 'optimal' supply depends upon the income distribution, as we have stated several times to date. A different distribution of income may give rise to different demands D_A and D_B, and therefore a different 'optimal' output.

> **Efficient supply of public goods** is where the marginal cost equals the sum of individual marginal valuations, and each individual consumes the same quantity.

Application Box 14.1: Are Wikipedia, Google and MOOCs public goods?

Wikipedia is one of the largest on-line sources of free information in the world. It is an encyclopedia that functions in multiple languages and that furnishes information on millions of topics. It is freely accessible, and is maintained and expanded by its users. Google is the most frequently used search engine on the World Wide Web. It provides information to millions of users simultaneously on every subject imaginable. But it is not quite free of charge; when the user searches she supplies information on herself that can be used profitably by Google in its advertising. MOOCs are 'monster open online courses' offered by numerous universities, frequently for no charge to the student. Are these services public goods in the sense we have described?

Very few goods and services are pure public goods, some have the major characteristics of public goods nonetheless. In this general sense, Google, Wikipedia and MOOCs have public good characteristics. Wikipedia is funded by philanthropic contributions, and its users expand its range by posting information on its servers. Google is funded from advertising revenue. MOOCs are funded by university budgets.

A pure public good is available to additional users at zero marginal cost. This condition is essentially met by these services since their server capacity rarely reaches its limit. Nonetheless, they are constantly adding server capacity, and in that sense cannot furnish their services to an unlimited number of additional users at no additional cost.

Knowledge is perhaps the ultimate public good; Wikipedia, Google and MOOCs all disseminate knowledge, knowledge which has been developed through the millennia by philosophers, scientists, artists, teachers, research laboratories and universities.

A challenge in providing the optimal amount of government-supplied public goods is to know the value that users may place upon them – how can the demand curves D_A and D_B, be ascertained, for example, in Figure 14.1? In contrast to markets for private goods, where consumer demands are essentially revealed through the process of purchase, the demands for public goods may have to be uncovered by means of surveys that are designed so as to elicit the true valuations that users place upon different amounts of a public good. A second challenge relates to the pricing and funding of public goods: For example, should highway lighting be funded from general tax revenue, or should drivers pay for it? These are complexities that are beyond our scope of our current inquiry.

Asymmetric information

Markets for information abound in the modern economy. Governments frequently supply information on account of its public good characteristics. But the problem of *asymmetric information* poses additional challenges. **Asymmetric information** is where at least one party in an economic relationship has less than full information. This situation characterizes many interactions: Life-insurance companies do not have perfect information on the lifestyle and health of their clients; used vehicle buyers may not know the history of the vehicles they are buying.

> **Asymmetric information** is where at least one party in an economic relationship has less than full information and has a different amount of information from another party.

Asymmetric information can lead to two kinds of problems. The first is **adverse selection**. For example, can the life-insurance company be sure that it is not insuring only the lives of people who are high risk and likely to die young? If primarily high-risk people buy such insurance then the insurance company must set its premiums accordingly: The company is getting an adverse selection rather than a random selection of clients. Frequently governments decide to run universal compulsory-membership insurance plans (auto or health are examples in Canada) precisely because they may not wish to charge higher rates to higher-risk individuals.

> **Adverse selection** occurs when incomplete or asymmetric information describes an economic relationship.

A related problem is **moral hazard**. If an individual does not face the full consequences of his actions, his behaviour may be influenced: if a homeowner has a fully insured home he may be less security conscious than an owner who does not.

In Chapter 7 we described how US mortgage providers lent large sums to borrowers with uncertain incomes in the early years of the new millennium. The individuals responsible for the lending were being rewarded on the basis of the amount lent, not the safety of the loan. Nor were the lenders responsible for loans that were not repaid. This ‘sub-prime mortgage crisis’ was certainly a case of moral hazard.

> **Moral hazard** may characterize behaviour where the costs of certain activities are not incurred by those undertaking them.

Solutions to these problems do not always involve the government, but in critical situations do. For example, the government requires most professional societies and orders to ensure that their members are trained, accredited and capable. Whether for a medical doctor, a plumber or an engineer, a license or certificate of competence is a signal that the work and advice of these professionals is *bona fide*. Equally, the government sets *standards* so that individuals do not have to incur the cost of ascertaining the quality of their purchases – bicycle helmets must satisfy specific crash norms; so too must air-bags in automobiles.

These situations differ from those where solutions to the information problem can be dealt with reasonably well in the market place. For example, with the advent of buyer and seller rating on *Airbnb*, a potential renter can learn of the quality of the accommodation he is considering, and the letor can assess the potential renter.

Concentration of power

Monopolistic and imperfectly-competitive market structures can give rise to inefficient outcomes, in the sense that the value placed on the last unit of output does not equal the cost at the margin. This arises because the supplier uses his market power in order to maximize profits by limiting output and selling at a higher price.

What can governments do about such power concentrations? Every developed economy has a body similar to Canada's *Competition Bureau*. Such regulatory bodies are charged with seeing that the interests of the consumer, and the economy more broadly, are represented in the market place. Interventions, regulatory procedures and efforts to prevent the abuse of market power come in a variety of forms. These measures are examined in Section 14.5.

Unfavourable market outcomes

Even if governments successfully address the problems posed by the market failures described above, there is nothing to guarantee that market-driven outcomes will be 'fair', or accord with the prevailing notions of justice or equity. The marketplace generates many low-paying jobs, unemployment and poverty. The concentration of economic power has led to the growth in income and wealth inequality in many economies. Governments, to varying degrees, attempt to moderate these outcomes through a variety of social programs and transfers that are discussed in Section 14.4.

14.2 Fiscal federalism: Taxing and spending

Canada is a federal state, in which the federal, provincial and municipal governments exercise different powers and responsibilities. In contrast, most European states are unitary and power is not devolved to their regions to the same degree as in Canada or the US or Australia. Federalism confers several advantages over a unitary form of government where an economy is geographically extensive, or where identifiable differences distinguish one region from another: Regions can adopt different policies in response to the expression of different preferences by their respective voters; smaller governments may be better at experimentation and the introduction of new policies than large governments; political representatives are 'closer' to their constituents.

Despite these advantages, the existence of an additional level of government creates a tension between these levels. Such tension is evident in every federation, and federal and provincial governments argue over the appropriate division of taxation powers and revenue-raising power in general. For example, how should the royalties and taxes from oil and gas deposits offshore be distributed – to the federal government or a provincial government?

In Canada, the federal government collects more in tax revenue than it expends on its own programs. This is a feature of most federations. The provinces simultaneously face a shortfall in their own revenues relative to their program expenditure requirements. The federal government therefore redistributes, or transfers, funds to the provinces so that the latter can perform their constitutionally-assigned roles in the economy. The fact that the federal government bridges this fiscal gap gives it a degree of power over the provinces. This influence is commonly termed **federal spending power**.

> **Spending power** of a federal government arises when the federal government can influence lower level governments due to its financial rather than constitutional power.

The principal revenue sources for financing federal government activity are given in Figure 14.2 for the fiscal year 2015-16, and the expenditure of these revenues is broken down in Figure 14.3. Further details are accessible in the Department of Finance's 'fiscal reference tables' at `https://www.fin.gc.ca/frt-trf/2019/frt-trf-19-eng.asp`. Total revenues for that fiscal year amounted to \$333.2b, and expenditures to \$346.2b.

Figure 14.2: Federal Government Revenues 2015–16

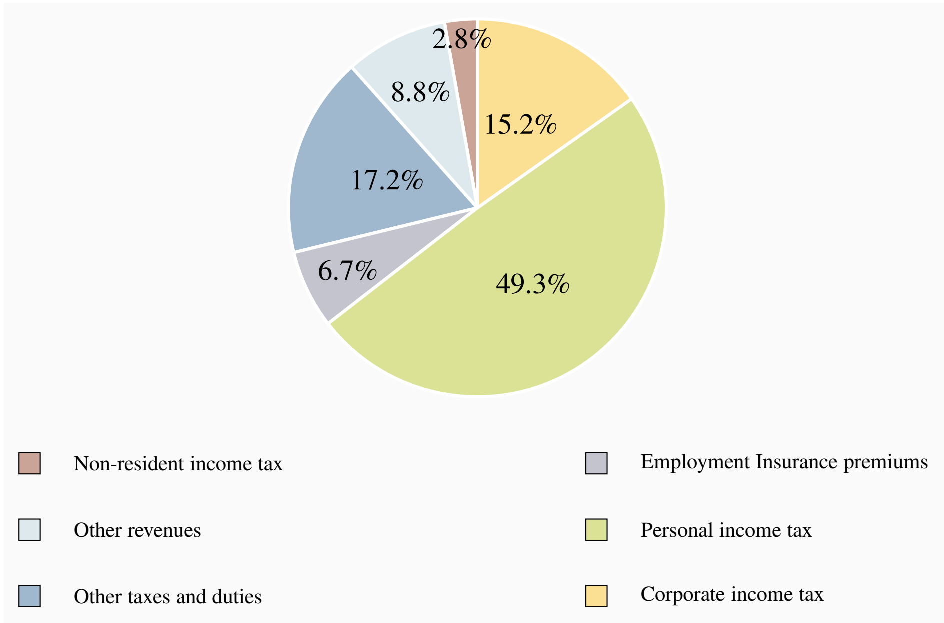
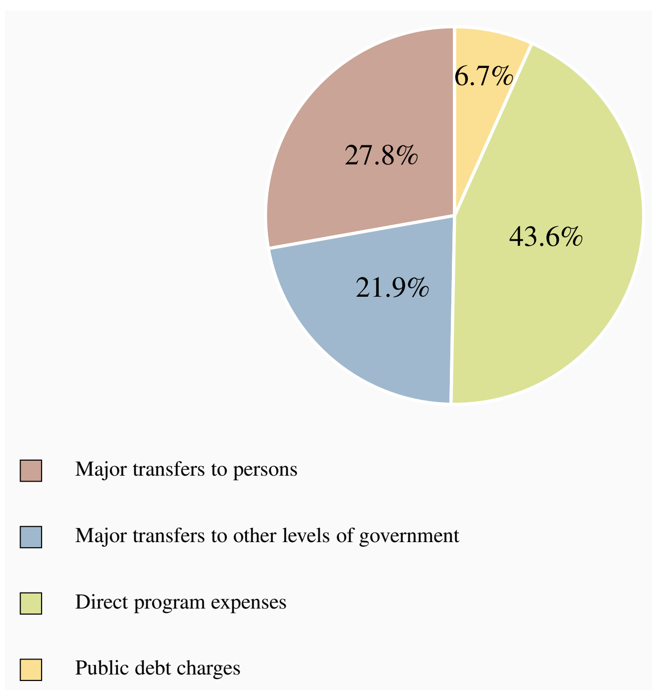


Figure 14.3: Federal expenditures 2015–16



The federal and provincial governments each transfer these revenues to individuals and other levels of government, supply goods and services directly, and also pay interest on accumulated borrowings – the national debt or provincial debt.

Provincial and local governments supply more goods and services than the federal government – health care, drug insurance, education and welfare are the responsibility of provincial and municipal governments. In contrast, national defence, the provision of main traffic arteries, Corrections Canada and a variety of transfer programs to individuals – such as Employment Insurance, Old Age Security and the Canada Pension Plan – are federally funded. The greater part of federal revenues goes towards *transfers* to individuals and provincial governments, as opposed to the supply of goods and services.

14.3 Federal-provincial fiscal relations

The federal government transfers revenue to the provinces using three main programs: Equalization, the Canada Social Transfer and the Canada Health Transfer. Each of these has a different objective. Equalization aims to reduce fiscal disparities among the provinces; The Canada Social Transfer (CST) is for educational and Social Assistance ('welfare') expenditures; The Canada Health Transfer (CHT) performs the same function for health.

Equalization and Territorial Funding

Canada's provinces receive unconditional funding through Canada's Equalization program, whereas the Territories receive federal funding through a separate mechanism - the Territorial Funding Formula.

> *"Parliament and the Government of Canada are committed to the principle of making equalization payments to ensure that provincial governments have sufficient revenues to provide reasonably comparable levels of public service at reasonably comparable levels of taxation."*

This statement, from Section 36(2) of the Constitution Act of 1982, defines the purpose of Equalization. Equalization payments are unconditional – receiving provinces are free to spend the funds on public services according to their own priorities, or even use the revenue to reduce their provincial taxes. Payments are calculated according to a formula that ensures those provinces with revenue-raising ability, or fiscal capacity, below a threshold or 'standard' receive payments from the federal government to bring their capacity up to that standard.

Equalization has gone through very many changes in the several decades of its existence. Its current rules and regulations reflect the 2006 recommendations of a federal Expert Panel. The fiscal capacity of a province is measured by its ability to raise revenues from five major sources: Personal and business income taxes, sales taxes, property taxes, and natural resources. This ability is then compared to the ability of all of the provinces combined to raise revenue; if a difference or shortfall exists, the federal government transfers revenue accordingly, with the amount determined by both the population of the province and the magnitude of its per-person shortfall. Data on annual transfers for Equalization, the Territorial Funding Formula, and the CHT and CST are available at `fin.gc.ca/fedprov/mtp-eng.asp`

The program transferred \$19.8b to the provinces in 2019-20. The recipiency status of some provinces varies from year to year. Variation in energy prices and energy-based government revenues are the principal cause of this. British Columbia, Alberta, Saskatchewan and Ontario tend to receive little or zero. Manitoba, Quebec and the Atlantic Provinces have been the major recipient provinces. Quebec receives the largest single amount – approximately two thirds of the total allocation, both on account of its population size and the fact that it has a lower than average fiscal capacity. The Territories received a total of \$3.9b in 2019-20 through the Territorial Funding Formula.

The Canada Social Transfer and the Canada Health Transfer

The CST is a block transfer to provinces in support of post-secondary education, Social Assistance and social services more generally. The CST came into effect in 2004. Prior to that date it was integrated with the health component of federal transfers in a program titled the Canada Health and Social Transfer (CHST). The objective of the separation was to increase the transparency and accountability of federal support for health while continuing to provide funding for other objectives. The CHT is the other part of the unbundled CHST: It provides funding to the provinces for their health expenditures.

The CST and CHT funding comes in two parts: A cash transfer and tax transfer. A tax transfer essentially provides the same support as a cash transfer of equal value, but comes in a different form. In 1977 the federal government agreed with provincial governments to reduce federal personal and corporate tax rates in order to permit the provincial governments to increase the corresponding provincial rates. The net effect was that the federal government got less tax revenue and the provinces got more. And to this day, the federal and provincial governments keep a record of the implied tax transfers that arise from this long-ago agreement. This is the tax transfer component of the CST and the CHT.

The CST support is allocated to provinces and territories on an equal per-capita basis to ensure equal support for all Canadians regardless of their place of residence. The CHT is distributed likewise, and it requires the provinces to abide by the federally-legislated *Canada Health Act*, which demands that provincial health coverage be comprehensive, universal, portable, accessible and publicly administered.

The CHT transfer amounted to \$40.4b and the CST amounted to \$14.6b, in cash, for the year 2019-20. Health care and health expenditures are a core issue of policy at all levels of government on account of the envisaged growth in these expenditures that will inevitably accompany the aging of the baby-boomers.

14.4 Government-to-individual transfers

Many Canadians take pride in Canada's extensive 'social safety net' that aims to protect individuals from misfortune and the reduction of income in old age. Others believe it is too generous. While it is more supportive than the safety net in the US, the Canadian safety net is no more protective than the nets of the developed economies of the European Union. The extent of such support depends in large measure upon the degree to which governments are willing to impose, and individuals are willing to pay, higher or lower tax rates. The major elements of this umbrella of programs are the following.

The *Canada and Quebec Pension Plans (C/QPP)* are funded from the contributions of workers and their employers. Contributions form 9.9% of an individual's earnings up to the maximum pensionable earnings (MPE) figure of $57,400 in 2019. The contributions are shared equally by employer and employee. The Canada and Quebec components of the plan operate similarly, but are managed separately. The contribution rate to the QPP stands at 10.65%. Contributions to the plans from workers and their employers are largely transferred immediately to retired workers. Part of the contributions is invested in a fund. The objective of the plans is to ensure that some income is saved for retirement. Many individuals are not very good at planning – they constantly postpone the decision to save, so the state steps in and requires them to save. An individual contributing throughout a full-time working lifecycle can expect an annual pension of about $14,000 in 2019. The Plans provide a maximum payout of 25% of maximum insurable earnings. The objective is to provide a minimum level of retirement income, not an income that will see individuals live in great comfort.

The C/QPP plans have contributed greatly to the reduction of poverty among the elderly since their introduction in the mid-sixties. The aging of the baby-boom generation – that very large cohort born in the late forties through to the early sixties – means that the percentage of the population in the post-65 age group has begun to increase. To meet this changing demographic, the federal and provincial governments reshaped the plans in the late nineties in order to put them on a sound financial footing – primarily by increasing contributions, that in turn will enable the build-up of a CPP 'fund' that will support the aged in the following decades.

A number of recent studies in Canada on the retirement savings practices of Canadians have proposed that households on average are not saving a sufficient amount for their retirement; many households may thus see a notable decline in their incomes upon retirement. In response to this finding, the federal government agreed with the provinces in June 2016, to add a supplement to the CPP. The new federal provisions, which will be phased in over the period 2019-2025, envisage an increase in contributions that will ultimately lead to a maximum replacement rate of 33% of MPE as opposed to the current goal of 25%. However, full benefits will be experienced only by individuals contributing for their complete lifecycle, meaning that full implementation will take about four decades.

Details of the CPP and the 2016 enhancements are to be found at `http://www.fin.gc.ca/n16/data/16-113_3-eng.asp`.

Old Age Security (OAS), the *Guaranteed Income Supplement (GIS)* and the *Spousal Allowance (SPA)* together form the second support leg for the retired. OAS is a payment made automatically to individuals once they attain the age of 65. The GIS is an additional payment made only to those on very low incomes – for example, individuals who have little income from their C/QPP or private pension plans. The SPA, which is payable to the spouse or survivor of an OAS recipient, accounts for a small part of the sums disbursed. As of 2019 the maximum annual OAS payment stood at

$7,360. The federal government in 2016 reversed a plan that would have seen the eligible age for receipt of OAS move to 67.

The payments for these plans come from the general tax revenues of the federal government. Unlike the C/QPP, the benefits received are not related to the contributions that an individual makes over the working lifecycle. This program has also had a substantial impact on poverty reduction among the elderly.

Employment Insurance (EI) and *Social Assistance (SA)* are designed to support, respectively, the unemployed and those with no other source of income. Welfare is the common term used to describe SA. Expenditures on EI and SA are strongly cyclical. At the trough of an economic cycle the real value of expenditures on these programs greatly exceeds expenditures at the peak of the cycle. Unemployment in Canada rose above 8% in 2009, and payments to the unemployed and those on welfare reflected this dire state. The strongly cyclical pattern of the cost of these programs reflects the importance of a healthy job market: Macroeconomic conditions have a major impact on social program expenditures.

EI is funded by contributions from employees and their employers. For each dollar contributed by the employee, the employer contributes $1.4. Premiums are paid on earned income up to a maximum insurable earnings (MIE) of $53,100 in 2019. The contribution rate for employees stood at 1.62% of MIE in 2017. EI contributions and pay-outs form part of the federal government's general revenues and expenditures. There is no separate 'fund' for this program. However it is expected to operate on a break-even basis over the longer term. To reflect this, the contribution rate fluctuates with a view to maintaining a balance between payouts and revenues over a seven-year planning period.

EI is called an insurance program, but in reality it is much more than that. Certain groups systematically use the program more than others – those in seasonal jobs, those in rural areas and those in the Atlantic Provinces, for example. Accordingly, using the terminology of Chapter 7, it is not everywhere an actuarially 'fair' insurance program. Benefits payable to unemployed individuals may also depend on their family size, in addition to their work history. While most payments go in the form of 'regular' benefits to unemployed individuals, the EI program also sponsors employee retraining, family benefits that cover maternity and paternity leave, and some other specific target programs for the unemployed.

Social Assistance is provided to individuals who are in serious need of financial support – having no income and few assets. Provincial governments administer SA, although the cost of the program is partly covered by federal transfers through the Canada Social Transfer. The nineteen nineties witnessed a substantial tightening of regulations virtually across the whole of Canada. Access to SA benefits is now more difficult, and benefits have fallen in real terms since the late nineteen eighties.

Welfare dependence peaked in Canada in 1994, when 3.1 million individuals were dependent upon support. As of 2019, the total is approximately half of this, on account of more stringent access conditions, reduced benefit levels and an improved job market. Some groups in Canada believe that benefits should be higher, others believe that making welfare too generous provides young individuals with the wrong incentives in life, and may lead them to neglect schooling and skill development.

Workers Compensation supports workers injured on the job. Worker/employer contributions and general tax revenue form the sources of program revenue, and the mix varies from province-to-province. In contrast to the macro-economy-induced swings in expenditures that characterize SA and EI since the early nineties, expenditures on Worker's Compensation have remained relatively constant.

Canada Child Benefit: The major remaining pillar in Canada's social safety net is the group of payments and tax credits aimed at supporting children: The Canada Child Tax Benefit (CCTB), the Universal Child Care Benefit and the National Child Benefit Supplement were repackaged in 2016 under the title Canada Child Benefit. Child support has evolved and been enriched over the last two decades, partly with the objective of reducing poverty among households with children, and partly with a view to helping parents receiving social assistance to transition back to the labour market. As of 2016, the federal government provides an annual payment to families with children. For each child under the age of 6 the payment is $6,639 and for each child aged 6-17 the payment is $5,602. Since these payment are primarily intended for households with low and middle incomes, the amounts are progressively clawed back once the household income reaches a threshold of $30,000.

Application Box 14.2: Government debts and deficits

Canada's expenditure and tax policies in the nineteen seventies and eighties led to the accumulation of large government debts, as a result of running fiscal deficits. By the mid-nineties the combined federal and provincial debt reached 100% of GDP, with the federal debt accounting for the larger share. This ratio was perilously high: Interest payments absorbed a large fraction of annual government revenues, which in turn limited the ability of the government to embark on new programs or enrich existing ones. Canada's debt rating on international financial markets declined.

In 1995, Finance Minister Paul Martin addressed this problem, and over the following years program spending was pared back. Ultimately, the economy expanded and by the end of the decade the annual deficits at the federal level were eliminated.

As of 2007 the ratio of combined federal and provincial debts stood at just over 60% of GDP. However, the Great Recession of 2008 and following years saw all levels of government experience deficits, with the result that this ratio of combined debt to GDP rose again. Growth in recent years has seen that ratio fall. As of 2018-19, federal interest payments on its debt account for about 7% of its revenues (this figure stood at 28% in the early nineties). At the time of writing, interest rates are low in developed economies and so the interest costs of government debt are low. Low borrowing costs are a reason why some people favor large government spending in the form of infrastructure projects. Those who are fiscally more conservative fear rising rates in the future. The recessionary impacts of the coronavirus pandemic of 2020 will add greatly to accumulated debt, particularly at the federal level.

Debts can be measured in more than a single manner. One measure of debt is the value of all federal government bonds and financial liabilities outstanding. As of 2019-20 this value was approximately $700b. In addition to this, the federal government has outstanding liabilities to the pensions of its retired employees, and not all of these liabilities have been covered by the contributions of those employees into their pension plans. The federal government also owns assets, both financial and physical - such as office buildings. Hence these assets offset the financial liabilities. To assess the total debt picture of the Canadian economy we need to add provincial and local government debts to the federal debts, and then consider the annual interest costs of this total. It turns out that the interest costs are just above 2% of GDP.

Source: Government of Canada, Fiscal Reference Tables: `https://www.fin.gc.ca/frt-trf/2019/frt-trf-19-eng.asp`

14.5 Regulation and competition policy

Goals and objectives

The goals of competition policy are relatively uniform across developed economies: The promotion of domestic competition; the development of new ideas, new products and new enterprises; the promotion of efficiency in the resource-allocation sense; the development of manufacturing and service industries that can compete internationally.

In addition to these economic objectives, governments and citizens frown upon monopolies or monopoly practices if they lead to an undue *concentration of political power*. Such power can lead to a concentration of wealth and influence in the hands of an elite.

Canada's regulatory body is the *Competition Bureau*, whose activity is governed primarily by the *Competition Act* of 1986. This act replaced the *Combines Investigation Act*. The *Competition Tribunal* acts as an adjudication body, and is composed of judges and non- judicial members. This tribunal can issue orders on the maintenance of competition in the marketplace. Canada has had anti-combines legislation since 1889, and the act of 1986 is the most recent form of such legislation and policy. The Competition Act does not forbid monopolies, but it does rule as unlawful the *abuse* of monopoly power. Canada's competition legislation is aimed at anti-competitive practices, and a full description of its activities is to be found on its website at www.competitionbureau.gc.ca. Let us examine some of these proscribed policies.

Anti-competitive practices

Anti-competitive practices may either limit entry into a sector of the economy or force existing competitors out. In either case they lead to a reduction in competition.

Mergers may turn competitive firms into a single organization with excessive market power. The customary justification for mergers is that they permit the merged firms to achieve scale economies that would otherwise be impossible. Such scale economies may in turn result in lower prices in the domestic or international market to the benefit of the consumer, but may alternatively reduce competition and result in higher prices. Equally important in this era of global competition is the impact of a merger on a firm's ability to compete internationally.

Mergers can be of the horizontal type (e.g. two manufacturers of pre-mixed concrete merge) or vertical type (a concrete manufacturer merges with a cement manufacturer). In a market with few suppliers mergers have the potential to reduce domestic competition.

Cartels aim to restrict output and thereby increase profits. These formations are almost universally illegal in individual national economies.

While cartels are one means of increasing prices, **price discrimination** is another, as we saw when studying monopoly behaviour. For example, if a concrete manufacturer makes her product available to large builders at a lower price than to small-scale builders – perhaps because the large builder has more bargaining power – then the small builder is at a competitive disadvantage in the construction business. If the small firm is forced out of the construction business as a consequence, then competition in this sector is reduced.

We introduced the concept of predatory pricing in Chapter 11. **Predatory pricing** is a practice that is aimed at driving out competition by artificially reducing the price of one product sold by a supplier. For example, a dominant nationwide transporter could reduce price on a particular route where competition comes from a strictly local competitor. By 'subsidizing' this route from profits on other routes, the dominant firm could undercut the local firm and drive it out of the market.

> **Predatory pricing** is a practice that is aimed at driving out competition by artificially reducing the price of one product sold by a supplier.

Suppliers may also **refuse to deal**. If the local supplier of pre-mixed concrete refuses to sell the product to a local construction firm, then the ability of such a downstream firm to operate and compete may be compromised. This practice is similar to that of **exclusive sales** and **tied sales**. An exclusive sale might involve a large vegetable wholesaler forcing her retail clients to buy *only* from this supplier. Such a practice might hurt the local grower of aubergines or zucchini, and also may prevent the retailer from obtaining *some* of her vegetables at a lower price or at a higher quality elsewhere. A **tied sale** is one where the purchaser must agree to purchase a *bundle* of goods from a supplier.

> **Refusal to deal**: an illegal practice where a supplier refuses to sell to a purchaser.
>
> **Exclusive sale**: where a retailer is obliged (perhaps illegally) to purchase all wholesale products from a single supplier only.

> **Tied sale**: one where the purchaser must agree to purchase a bundle of goods from a supplier.

Resale price maintenance involves the producer requiring a retailer to sell a product at a specified price. This practice can hurt consumers since they cannot 'shop around'. In Canada, we frequently encounter a 'manufacturer's suggested retail price' for autos and durable goods. But since these prices are not *required*, the practice conforms to the law.

> **Resale price maintenance** is an illegal practice wherein a producer requires sellers to maintain a specified price.

Bid rigging is an illegal practice in which normally competitive bidders conspire to fix the awarding of contracts or sales. For example, two builders, who consider bidding on construction projects, may decide that one will bid seriously for project X and the other will bid seriously on project Y. In this way they conspire to reduce competition in order to make more profit.

> **Bid rigging** is an illegal practice in which bidders (buyers) conspire to set prices in their own interest.

Deception and dishonesty in promoting products can either short-change the consumer or give one supplier an unfair advantage over other suppliers.

Enforcement

The Competition Act is enforced through the Competition Bureau in a variety of ways. Decisions on acceptable business practices are frequently reached through study and letters of agreement between the Bureau and businesses. In some cases, where laws appear to have been violated, criminal proceedings may follow.

Regulation, deregulation and privatization

The last three decades have witnessed a significant degree of privatization and deregulation in Canada, most notably in the transportation, communication and energy sectors. Modern deregulation in the US began with the passage of the *Airline Deregulation Act* of 1978, and was pursued with great energy under the Reagan administration in the eighties. The Economic Council of Canada produced an influential report in 1981, titled "Reforming Regulation," on the impact of regulation and possible deregulation of specific sectors. The Economic Council proposed that regulation in some sectors was inhibiting competition, entry and innovation. As a consequence, the interests of the consumer were in danger of becoming secondary to the interests of the suppliers.

Telecommunications provision, in the era when the telephone was the main form of such communication, was traditionally viewed as a natural monopoly. The Canadian Radio and Telecommunications Commission (CRTC) regulated its rates. The industry has developed dramatically in the last two decades with the introduction of satellite-facilitated communication, the internet, multi-purpose cable networks, cell phones and service integration.

Transportation, in virtually all forms, has been deregulated in Canada since the nineteen eighties. Railways were originally required to subsidize the transportation of grain under the *Crow's Nest Pass* rate structure. But the subsidization of particular markets requires an excessive rate elsewhere, and if the latter markets become subject to competition then a competitive system cannot function. This structure, along with many other anomalies, was changed with the passage of the *Canada Transportation Act* in 1996.

Trucking, historically, has been regulated by individual provinces. Entry was heavily controlled prior to the federal *National Transportation Act* of 1987, and subsequent legislation introduced by a number of provinces, have made for easier entry and a more competitive rate structure.

Deregulation of the airline industry in the US in the late seventies had a considerable influence on thinking and practice in Canada. The Economic Council report of 1981 recommended in favour of easier entry and greater fare competition. These policies were reflected in the 1987 National Transportation Act. Most economists are favourable to deregulation and freedom to enter, and the US experience indicated that cost reductions and increased efficiency could follow. In 1995 an agreement was reached between the US and Canada that provided full freedom for Canadian carriers to move passengers to any US city, and freedom for US carriers to do likewise, subject to a phase-in provision.

The National Energy Board regulates the development and transmission of oil and natural gas. But earlier powers of the Board, involving the regulation of product prices, were eliminated in 1986, and controls on oil exports were also eliminated.

Agriculture remains a highly controlled area of the economy. Supply 'management', which is really supply restriction, and therefore 'price maintenance', characterizes grain, dairy, poultry and other products. Management is primarily through provincial marketing boards.

The role of the sharing economy

The arrival of universal access to the internet has seen the emergence of what is known as the **Sharing economy** throughout the world. This expression is used to describe commercial activities that, in the first place, are internet-based. Second, suppliers in the sharing economy use resources in the market place that were initially aimed at a different purpose. *Airbnb* and *Uber* are good examples of companies in sectors of the economy where sharing is possible. In Uber's case, the 'ride-share' drivers initially purchased their vehicles for private use, and subsequently redirected them to commercial use. *Airbnb* is a communication corporation that enables the owners of spare home capacity to sell the use of that capacity to short-term renters. With the maturation of such corporations, the concept of 'initial' and 'secondary' use becomes blurred.

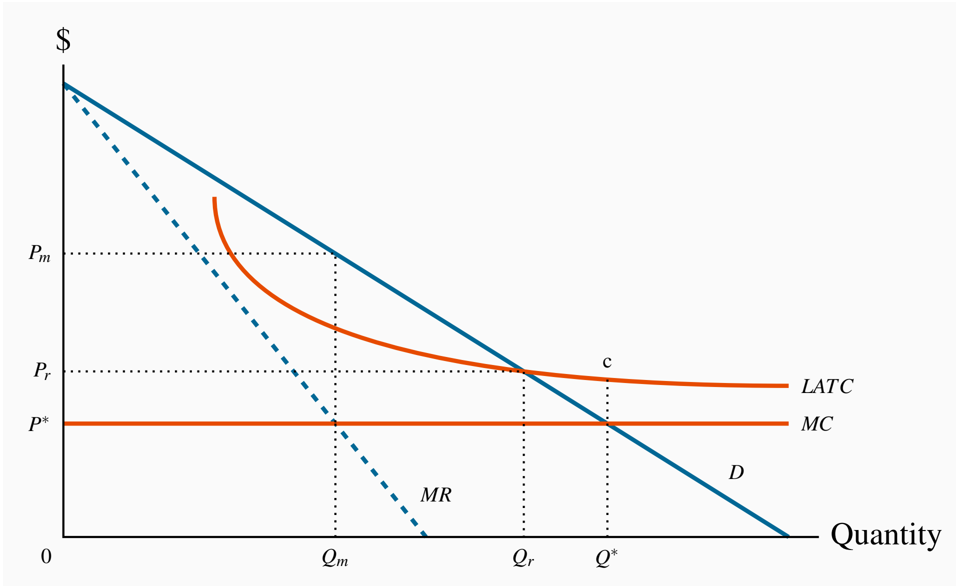
> **Sharing economy**: involves enterprises that are internet based, and that use production resources that have use outside of the marketplace.

The importance of the sharing economy is that it provides an additional source of competition to established suppliers, and therefore limits the market power of the latter. At the same time, the emergence of the sharing economy poses a new set of regulatory challenges: If traditional taxis are required to purchase operating permits (medallions), and the ride-share drivers do not require such permits, is there a reasonable degree of competition in the market, and if not what is the appropriate solution? Should the medallion requirement be abolished, or should ride-share drivers be required to purchase one? In the case of *Airbnb*, the suppliers operate outside of the traditional 'hotel' market. In general they do not charge sales taxes or face any union labour agreements. What is the appropriate response from governments? And how should the sharing economy be taxed?

Price regulation

Regulating monopolistic sectors of the economy is one means of reducing their market power. In Chapter 11 it was proposed that indefinitely decreasing production costs in an industry means that the industry might be considered as a 'natural' monopoly: Higher output can be produced at lower cost with fewer firms. Hence, a single supplier has the potential to supply the market at a lower unit cost; unless, that is, such a single supplier uses his monopoly power. To illustrate how the consumer side may benefit from this production structure through regulation, consider Figure 14.4. For simplicity suppose that long-run marginal costs are constant and that average costs are downward sloping due to an initial fixed cost. The profit-maximizing (monopoly) output is where $MR = MC$ at Q_m and is sold at the price P_m. This output is inefficient because the willingness of buyers to pay for additional units of output exceeds the additional cost. On this criterion the efficient output is Q^*. But $LATC$ exceeds price at Q^*, and therefore it is not feasible for a producer.

Figure 14.4: Regulating a decreasing-cost supplier



The profit-maximizing output is Q_m, where $MR = MC$ and price is P_m. This output is inefficient because marginal benefit is greater than MC. Q^* is the efficient output, but results in losses because $LATC > P$ at that output. A regulated price that covers costs is where $LATC = DQ_r$. This is closer to the efficient output Q^* than the monopoly output Q_m.

One solution is for the regulating body to set a price-quantity combination of P_r, and Q_r, where price equals average cost and therefore generates a normal rate of profit. This output level is still lower than the efficient output level Q^*, but is more efficient than the profit-maximizing output Q_m. It is more efficient in the sense that it is closer to the efficient output Q^*. A problem with such a strategy is that it may induce lax management: If producers are allowed to charge an average-cost price, then there is a reduced incentive for them to keep strict control of their costs in the absence of competition in the marketplace.

A second solution to the declining average cost phenomenon is to implement what is called a **two-part tariff**. This means that customers pay an 'entry fee' in order to be able to purchase the good. For example, in many jurisdictions hydro or natural gas subscribers may pay a fixed charge per month for their supply line and supply guarantee, and then pay an additional charge that varies with quantity. In this way it is possible for the supplier to charge a price per unit of output that is closer to marginal cost and still make a profit, than under an average cost pricing formula. In terms of Figure 14.4, the total value of entry fees, or fixed components of the pricing, would have to cover the difference between MC and $LATC$ times the output supplied. In Figure 14.4 this implies that if the efficient output Q^* is purchased at a price equal to the MC the producer loses the amount $(c - MC)$ on each unit sold. The access fees would therefore have to cover at least this value.

Such a solution is appropriate when fixed costs are high and marginal costs are low. This situation is particularly relevant in the modern market for telecommunications: The cost to suppliers of marginal access to their networks, whether it be for internet, phone or TV, is negligible compared to the cost of maintaining the network and installing capacity.

> **Two-part tariff**: involves an access fee and a per unit of quantity fee.

Finally, a word of caution: Nobel Laureate George Stigler has argued that there is a danger of regulators becoming too close to the regulated, and that the relationship can evolve to a point where the regulator may protect the regulated firms. In contrast, Professor Philippon of New York University argues that regulators are not regulating sufficiently in the US: they have permitted an excessive number of mergers that have, in turn, reduced competition.

Key Terms

Market failure defines outcomes in which the allocation of resources is not efficient.

Public goods are non-rivalrous, in that they can be consumed simultaneously by more than one individual; additionally they may have a non-excludability characteristic.

Efficient supply of public goods is where the marginal cost equals the sum of individual marginal valuations, and each individual consumes the same quantity.

Asymmetric information is where at least one party in an economic relationship has less than full information and has a different amount of information from another party.

Adverse selection occurs when incomplete or asymmetric information describes an economic relationship.

Moral hazard may characterize behaviour where the costs of certain activities are not incurred by those undertaking them.

Spending power of a federal government arises when the federal government can influence lower level governments due to its financial rather than constitutional power.

Predatory pricing is a practice that is aimed at driving out competition by artificially reducing the price of one product sold by a supplier.

Refusal to deal: an illegal practice where a supplier refuses to sell to a purchaser.

Exclusive sale: where a retailer is obliged (perhaps illegally) to purchase all wholesale products from a single supplier only.

Tied sale: one where the purchaser must agree to purchase a bundle of goods from the one supplier.

Resale price maintenance is an illegal practice wherein a producer requires sellers to maintain a specified price.

Bid rigging is an illegal practice in which bidders (buyers) conspire to set prices in their own interest.

Sharing economy: involves enterprises that are internet based, and that use production resources that have use outside of the marketplace.

Two-part tariff: involves an access fee and a per unit of quantity fee.

Exercises for Chapter 14

Exercise 14.1 An economy is composed of two individuals, whose demands for a public good – street lighting – are given by $P = 12 - (1/2)Q$ and $P = 8 - (1/3)Q$.

(a) Graph these demands on a diagram, for values of $Q = 1, \ldots, 24$.

(b) Graph the total demand for this public good by summing the demands vertically, specifying the numerical value of each intercept.

(c) Let the marginal cost of providing the good be \$5 per unit. Illustrate graphically the efficient supply of the public good (Q^*) in this economy.

(d) Illustrate graphically the area that represents the total value to the consumers of the amount Q^*.

Exercise 14.2 In Exercise 14.1, suppose a new citizen joins the economy, and her demand for the public good is given by $P = 10 - (5/12)Q$.

(a) Add this individual's demand curve to the graphic for the above question and graph the new total demand curve, specifying the intercept values.

(b) Illustrate the area on your graph that represents the new total value to the three citizens of the optimal amount supplied.

(c) Illustrate graphically the net value to society of the new Q^* – the total value minus the total cost.

Exercise 14.3 An industry that is characterized by a decreasing cost structure has a demand curve given by $P = 100 - Q$ and the marginal revenue curve by $MR = 100 - 2Q$. The marginal cost is $MC = 4$, and average cost is $AC = 4 + 188/Q$.

(a) Graph this cost and demand structure. [*Hint*: This graph is similar to Figure 14.4.]

(b) Illustrate the efficient output and the monopoly output for the industry.

(c) Illustrate on the graph the price the monopolist would charge if he were unregulated.

Exercise 14.4 *Optional*: In Question 14.3, suppose the government decides to regulate the behaviour of the supplier, in the interests of the consumer.

(a) Illustrate graphically the price and output that would emerge if the supplier were regulated so that his allowable price equalled average cost.

(b) Is this greater or less than the efficient output?

(c) Compute the AC and P that would be charged with this regulation.

(d) Illustrate graphically the deadweight loss associated with the regulated price and compare it with the deadweight loss under monopoly.

Exercise 14.5 *Optional*: As an alternative to regulating the supplier such that price covers average total cost, suppose that a two part tariff were used to generate revenue. This scheme involves charging the *MC* for each unit that is purchased and in addition charging each buyer in the market a fixed cost that is independent of the amount he purchases. If an efficient output is supplied in the market, illustrate graphically the total revenue to be obtained from the component covering a price per unit of the good supplied, and the component covering fixed cost.

Chapter 15

International trade

In this chapter we will explore:

15.1 Trade in our daily lives

Virtually every economy in the modern world trades with other economies – they are what we call 'open' economies. Evidence of such openness is everywhere evident in our daily life. The world eats Canadian wheat; China exports manufactured goods to almost anywhere we can think of; and Canadians take their holidays in Florida.

As consumers we value the choice and variety of products that trade offers. We benefit from lower prices than would prevail in a world of protectionism. At the same time there is a constant chorus of voices calling for protection from international competition: Manufacturers are threatened by production in Asia; farmers fight against the imports of poultry, beef, and dairy products; even the service sector is concerned about offshore competition from call centres and designers. In this world of competing views it is vital to understand how trade has the potential to improve the well-being of economies.

This chapter examines the theory of international trade, trade flows, and trade policy: Who trades with whom, in what commodities, and why. In general, countries trade with one another because they can buy foreign products at a lower price than it costs to make them at home. International trade reflects specialization and exchange, which in turn improve living standards. It is cost differences between countries rather than technological differences that drive trade: In principle, Canada could supply Toronto with olives and oranges grown in Nunavut greenhouses, but it makes more sense to import them from Greece, Florida or Mexico.

Trade between Canada and other countries differs from trade between provinces. By definition, international trade involves jumping a border, whereas most trade within Canada does not. Internal borders are present in some instances – for example when it comes to recognizing professional qualifications acquired out-of-province. In the second instance, international trade may involve different currencies. When Canadians trade with Europeans the trade is accompanied by financial transactions involving Canadian dollars and Euros. A Canadian buyer of French wine pays in Canadian dollars, but the French vineyard worker is paid in euros. Exchange rates are one factor in determining national competitiveness in international markets. Evidently, not every international trade requires currency trades at the same time – most members of the European Union use the Euro. Indeed a common currency was seen as a means of facilitating trade between member nations of the EU, and thus a means of integrating the constituent economies more effectively.

It is important at the outset to emphasize that while trade has the potential to improve aggregate well being in the trading economies, this does not mean that every citizen will benefit; some will gain others will lose out. Buyers usually gain as a result of having a wider array of products to purchase at lower prices. Successful producers may benefit from production efficiencies associated with accessing a global supply chain, while others may be squeezed by international competition. The former may export more and employ more workers, the latter may contract and lay off employees.

15.2 Canada in the world economy

World trade has grown rapidly since the end of World War II, indicating that trade has become ever more important to national economies. Canada has been no exception. Canada signed the Free Trade Agreement with the US in 1989, and this agreement was expanded in 1994 when Mexico was included under the North America Free Trade Agreement (NAFTA). Imports and exports rose dramatically, from approximately one quarter to forty percent of GDP. Canada is now what is termed a very 'open' economy – one where trade forms a large fraction of total production. In early 2017, Canada and the EU signed a trade agreement - the Comprehensive Economic and Trade Agreement (CETA). Under this agreement tariffs will be phased out or reduced in most areas of trade over a several-year period. Canada also signed the Comprehensive and Progressive Agreement on Trans-Pacific Trade in 2018 that has the same objective of reducing trade barriers between 11 Pacific-Rim member states.

Smaller economies are typically more open than large economies—Belgium and the Netherlands depend upon trade more than the United States. This is because large economies tend to have a sufficient variety of resources to supply much of an individual country's needs. The European Union is similar, in population terms, to the United States, but it is composed of many distinct economies. Some European economies are equal in size to individual American states. But trade between California and New York is not international, whereas trade between Italy and the Spain is.

Because our economy is increasingly open to international trade, events in the world economy affect our daily lives much more than in the past. The conditions in international markets for basic commodities and energy affect all nations, both importers and exporters. For example, the prices of primary commodities on world markets increased dramatically in the latter part of the 2000s.

Higher prices for grains, oil, and fertilizers on world markets brought enormous benefits to Canada, particularly the Western provinces, which produce these commodities. In contrast, by early 2015, many of these prices dropped dramatically and Canadian producers suffered as a consequence.

The service sector accounts for more of our GDP than the manufacturing sector. As incomes grow, the demand for health, education, leisure, financial services, tourism, etc., dominates the demand for physical products. Technically, the income elasticity demand for the former group exceeds the income elasticity of demand for the latter. Internationally, while trade in services is growing rapidly, it still forms a relatively small part of total world trade. Trade in goods—merchandise trade—remains dominant, partly because many countries import unfinished goods, add some value, and re-export them. Even though the value added from such import-export activity may make just a small contribution to GDP, the gross flows of imports and exports can still be large relative to GDP. The transition from agriculture to manufacturing and then to services has been underway in developed economies for over a century. This transition has been facilitated in recent decades by the communications revolution and globalization. Globalization has seen a rapid shift in merchandise production from the developed to the developing world.

Table 15.1 shows the patterns of Canadian merchandise trade in 2018. The US is Canada's major trading partner, buying almost three quarters of our exports and supplying almost two thirds of imports. Table 15.2 details exports and imports by type. Although exports of resource-based products account for only about 40 percent of total merchandise exports, Canada is still viewed as a resource-based economy. This is in part because manufactures account for almost 80 percent of US and European merchandise exports and about 60 percent of Canadian exports. Nevertheless, Canada has important strength in machinery, equipment, and automotive products.

Table 15.1: Canada's Merchandise Trade Patterns 2018

Country	**Exports to**	**Imports from**
United States	73.9	64.4
European Union	7.9	10.5
China	5.0	7.6
Mexico	1.6	3.4
Others	11.6	14.1
Total	100	100
Dollar Total	585,255.7	607,205.4

Source: Adapted from Statistics Canada Table 12-10-0011-01

Table 15.2: Canadian Trade by Merchandise Type 2017

Sector	Exports	Imports
Farm, fishing, and intermediate food products	6.5	3.0
Energy products	20.5	5.5
Metal ores and non-metallic minerals	3.8	2.3
Metal and non-metallic mineral products	11.9	7.6
Basic and industrial chemical, plastic and rubber products	6.6	8.5
Forestry products and building and packaging materials	8.5	4.4
Industrial machinery, equipment and parts	5.3	9.7
Electronic and electrical equipment and parts	3.5	11.8
Motor vehicles and parts	16.3	20.0
Aircraft and other transportation equipment and parts	3.7	3.7
Consumer Goods	12.3	21.9
Special transactions trade	1.1	1.6
Total	100	100
Total dollar value in millions	500,892.6	561,425.9

Source: Adapted from Statistics Canada Table 12-10-0002-01

15.3 The gains from trade: Comparative advantage

In the opening chapter of this text we emphasized the importance of opportunity cost and differing efficiencies in the production process as a means of generating benefits to individuals through trade in the marketplace. The simple example we developed illustrated that, where individuals differ in their efficiency levels, benefits can accrue to each individual as a result of specializing and trading. In that example it was assumed that individual A had an absolute advantage in producing one product and that individual Z had an absolute advantage in producing the second good. This set-up could equally well be applied to two economies that have different efficiencies and are considering trade, with the objective of increasing their consumption possibilities. Technically, we could replace Amanda and Zoe with Argentina and Zambia, and nothing in the analysis would have to change in order to illustrate that consumption gains could be attained by both Argentina and Zambia as a result of specialization and trade.

Remember: The opportunity cost of a good is the quantity of another good or service given up in order to have one more unit of the good in question.

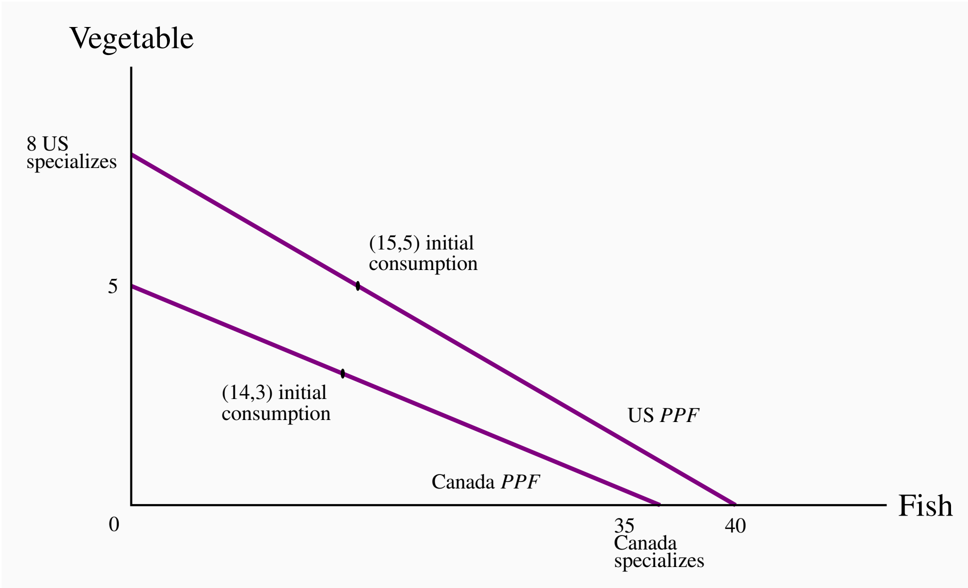
So, let us now consider two economies with differing production capabilities, as illustrated in Figures 15.1 and 15.2. In this instance it is assumed that one economy has an absolute advantage in both goods, but the degree of that advantage is greater in one good than the other. In international trade language, there exists a comparative advantage as well as an absolute advantage. It is frequently a surprise to students that this situation has the capacity to yield consumption advantages to each economy, even though one is absolutely more efficient in producing both of the goods. This is termed the **principle of comparative advantage**, and it states that even if one country has an absolute advantage in producing both goods, gains to specialization and trade still materialize, provided the opportunity cost of producing the goods differs between economies. This is a remarkable result, and much less intuitive than the principle of absolute advantage. We explore it with the help of the example developed in Figures 15.1 and 15.2.

> **Principle of comparative advantage** states that even if one country has an absolute advantage in producing both goods, gains to specialization and trade still materialize, provided the opportunity cost of producing the goods differs between economies.

We will name these two imaginary economies the US and Canada. Their production possibilities are defined by the PPFs in Figure 15.1. Canada can produce 5 units of V or 35 units of F, or any combination defined by the line joining these points. With the same resources the US can produce $8V$ or $40F$, or any combination defined by its PPF[1]. With no trade, Canadians and Americans consume a combination of the goods defined by some point on their respective PPFs. The opportunity cost of a unit of V in Canada is $7F$ (the slope of Canada's PPF is $5/35 = 1/7$). In the US the opportunity cost of one unit of V is $5F$ (slope is $8/40 = 1/5$). In this set-up the US is more efficient in producing V than F relative to Canada, as reflected by the opportunity costs. Hence we say that the *US has a comparative advantage in the production of* V <u>and</u> that *Canada has therefore a comparative advantage in producing* F.

[1]Note that we are considering the PPFs to be straight lines rather than concave shapes. The result we illustrate here carries over to that case also, but it is simpler to illustrate with the linear PPFs.

Figure 15.1: Comparative advantage – production



Canada specializes completely in Fish at 35, where it has a comparative advantage. Similarly, the US specializes in Vegetable at 8. They trade at a rate of 1 : 6. The US trades $3V$ to Canada in return for $18F$.

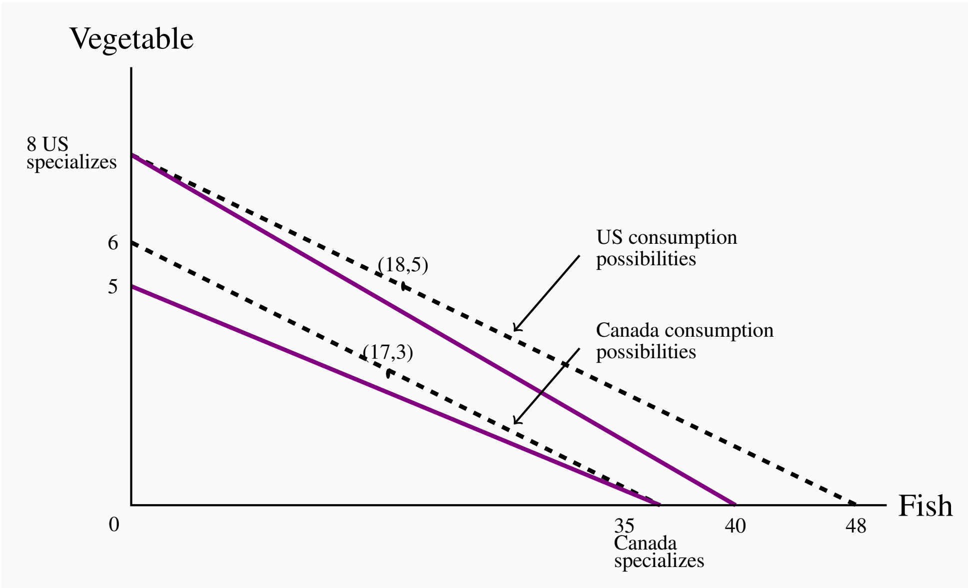
Prior to trade each economy is producing all of the goods it consumes. This no-trade state is termed **autarky**.

> **Autarky** denotes the no-trade situation.

The gains from trade

We now permit each economy to specialize in producing where it has a comparative advantage. So Canada specializes completely by producing $35F$ and the US produces $8V$. Having done this the economies must now agree on the terms of trade. The **terms of trade** define the rate at which the two goods will trade post-specialization. Let us suppose that a bargaining process leads to agreement that one unit of V will trade for six units of F. Such a trading rate, one that lies between the opportunity costs of each economy, benefits both economies. This exchange rate lies between Canada's opportunity cost of 1 : 7 and the US opportunity cost of 1 : 5. By specializing in F, Canada can now obtain an additional unit of V by sacrificing six units of F, whereas pre-trade it had to sacrifice seven units of F for a unit of V. Technically, by specializing in F and trading at a rate of 1 : 6 Canada's consumption possibilities have expanded and are given by the consumption possibility frontier (CPF) illustrated in Figure 15.2. The **consumption possibility frontier** defines what an economy can consume after production specialization and trade.

Figure 15.2: Comparative advantage – consumption



Post specialization the economies trade $1V$ for $6F$. Total production is $35F$ plus $8V$. Hence one consumption possibility would be (18,5) for the US and (17,3) for Canada. Here Canada exchanges $18F$ in return for $3V$.

The US also experiences an improved set of consumption possibilities. By specializing in V and trading at a rate of 1 : 6 its CPF lies outside its PPF and this enables it to consume more than in the pre-specialization state, where its CPF was defined by its PPF.

Evidently, the US and Canada CPFs are parallel since they trade with each other at the same rate: If Canada exports six units of F for every unit of V that it imports from the US, then the US must import the same six units of F for each unit of V it exports to Canada. The remarkable outcome here is that, even though one economy is more efficient in producing each good, specialization still leads to gains for both economies. The gain is illustrated by the fact that each economy's consumption possibilities lie outside of its production possibilities[2].

> **Terms of trade** define the rate at which the goods trade internationally.
>
> **Consumption possibility frontier** defines what an economy can consume after production specialization and trade.

[2] To illustrate the gains numerically, let Canada import $3V$ from the US in return for exporting $18F$. Note that this is a trading rate of 1 : 6. Hence, Canada consumes $3V$ and $17F$ (Canada produced $35F$ and exported $18F$, leaving it with $17F$). It follows that the US consumes $5V$, having exported $3V$ of the $8V$ it produced, and obtained in return $18F$ in imports. The new consumption bundles are illustrated in the figure: $(17,3)$ for Canada and $(18,5)$ for the US. These consumption bundles clearly represent an improvement over the autarky situation. For example, if Canada had wished to consume $3V$ pre-trade, it would only have been able to consume $14F$, whereas with trade it can consume $17F$.

Comparative advantage and factor endowments

A traditional statement of why comparative advantage arises is that economies have different endowments of the factors of production – land, capital and labour endowments differ. A land endowment that facilitates the harvesting of grain (Saskatchewan) or the growing of fruit (California) may be innate to an economy. We say that wheat production is *land intensive*, that aluminum production is *power intensive*, that research and development is *skill intensive*, that auto manufacture is *capital intensive,* that apparel is *labour intensive*. Consequently, if a country is well endowed with some particular factors of production, it is to be expected that it will specialize in producing goods that use those inputs. A relatively abundant supply or endowment of one factor of production tends to make the cost of using that factor relatively cheap: It is relatively less expensive to produce clothing in China and wheat in Canada than the other way around. This explains why Canada's Prairies produce wheat, why Quebec produces aluminum, why Asia produces apparel. But endowments can evolve.

How can we explain why Switzerland specializes in watches, precision instruments, and medical equipment, while Vietnam specializes in rice, tourism and manufactured goods and components? Evidently, Switzerland made a decision to educate its population and invest in the capital required to produce these goods. It was not naturally endowed with these skills, in the same way that Greece is endowed with sun or Saskatchewan is endowed with fertile flat land.

While we have demonstrated the principle of comparative advantage using a two-good example (since we are constrained by the geometry of two dimensions), the conclusions carry over to the case of many goods. Furthermore, the principle has many applications. For example, if one person in the household is more efficient at doing all household chores than another, there are still gains to specialization provided the efficiency differences are not all identical. This is the principle of comparative advantage at work in a microcosm.

Application Box 15.1: The one hundred mile diet

In 2005 two young British Columbians embarked on what has famously become known as the 'one hundred mile diet'—a challenge to eat and drink only products grown within this distance of their home. They succeeded in doing this for a whole year, wrote a book on their experience and went on to produce a TV series. They were convinced that such a project is good for humanity, partly because they wrapped up ideas on organic farming and environmentally friendly practices in the same message.

Reflect now on the implications of this superficially attractive program: If North Americans were to espouse this diet, it would effectively result in the closing down of the midwest of the Continent. From Saskatchewan to Kansas, we are endowed with grain-producing land that is the envy of the planet. But since most of this terrain is not within 100 miles of any big cities, these deluded advocates are proposing that we close up the production of grains and cereals exactly in those locations where such production is extraordinarily efficient. Should we sacrifice grains and cereals completely in this hemisphere, or just cultivate them on a hillside close to home, even if the resulting cultivation were to be more labour and fuel intensive? Should we produce olives in greenhouses in Edmonton rather than importing them from the Mediterranean, or simply stop eating them? Should we sacrifice wine and beer in North Battleford because insufficient grapes and hops are grown locally?

Would production in temperate climates really save more energy than the current practice of shipping vegetables and fruits from a distance—particularly when there are returns to scale associated with their distribution? The 'one hundred mile diet' is based on precepts that are contrary to the norms of the gains from trade. In its extreme the philosophy proposes that food exports be halted and that the world's great natural endowments of land, water, and sun be allowed to lie fallow. Where would that leave a hungry world?

Table 15.1 shows the patterns of Canadian merchandise trade in 2008. The United States was and still is Canada's major trading partner, buying almost three quarters of our exports and supplying almost two thirds of Canadian imports. Table 15.2 details exports by type. Although exports of resource-based products account for only about 40 percent of total exports, Canada is now viewed as a resource-based economy. This is in part because manufactured products account for almost 80 percent of US and European exports but only about 60 percent of Canadian exports. Nevertheless, Canada has important export strength in machinery, equipment, and automotive products.

15.4 Returns to scale and dynamic gains from trade

The theory of comparative advantage explains why economies should wish to trade. The theory is based upon the view that economies are 'inherently' different in their production capabilities. But trade is influenced by more than these differences. We will explore how returns to scale may be exploited to generate benefits from trade, and also how economies might gain from one-another by learning as a result of trading. This learning can increase domestic productivity.

Returns to scale

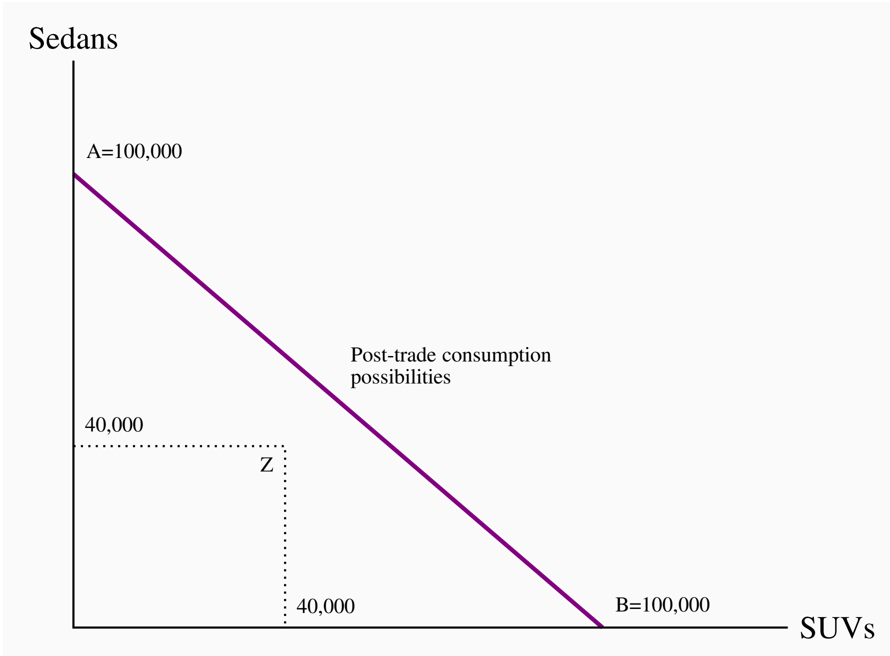
One of the reasons Canada signed the North America Free trade Agreement (NAFTA) was that economists convinced the Canadian government that a larger market would enable Canadian producers to be even *more efficient* than in the presence of trade barriers. Rather than opening up trade in order to take advantage of existing comparative advantage, it was proposed that efficiencies would actually increase with market size. This argument is easily understood in terms of increasing returns to scale concepts that we developed in Chapter 8. Essentially, economists suggested that there were several sectors of the Canadian economy that were operating on the downward sloping section of their long-run average cost curve.

Increasing returns are evident in the world market place as well as the domestic marketplace. Witness the small number of aircraft manufacturers—*Airbus* and *Boeing* are the world's two major manufacturers of large aircraft. Enormous fixed costs—in the form of research, design, and development—or capital outlays frequently result in decreasing unit costs, and the world marketplace can be supplied at a lower cost if some specialization can take place. *Facebook* is the giant in social media. Entertainment streaming companies like *Netflix*, *Amazon*, *Disney* and *Hulu* are few in number because of scale economies. Consider the specific example of automotive trade. In North America, Canadian auto plants produce different vehicle models than their counterparts in the US. Canada exports some *models* of a given manufacturer to the United States and imports other *models*. This is the phenomenon of **intra-industry trade** and **intra-firm trade**. How can we explain these patterns?

> **Intra-industry trade** is two-way international trade in products produced within the same industry.
>
> **Intra-firm trade** is two-way trade in international products produced within the same firm.

In the first instance, intra-industry trade reflects the preference of consumers for a choice of brands; consumers do not all want the same vehicle, or the same software, or the same furnishings. The second element to intra-industry trade is that increasing returns to scale characterize many production processes. Let us see if we can transform the returns to scale ideas developed in earlier chapters into a production possibility framework.

Figure 15.3: Intra industry trade



Hunda can produce either 100,000 of each vehicle or 40,000 of both in each plant. Hence production possibilities are given by the points A, Z, and B. Pre-trade it produces at Z in each economy due to trade barriers. Post-trade it produces at A in one economy and B in the other, and ships the vehicles internationally. Total production increases from 160,000 to 200,000 using the same resources.

Consider the example presented in Figure 15.3, where the hypothetical company *Hunda Motor Corporation* currently has a large assembly plant in *each of* Canada and the US where it produces two types of vehicles; sedans and sports utility vehicles (SUVs). Intially, restrictions on trade in automobiles, in the form of tariffs, between the two countries make it too costly to ship models across the border. Hence Hunda produces both sedans and SUVs in each plant. But for several reasons, *switching between model production is costly and results in reduced output*. Hunda can produce 40,000 vehicles of each type per annum in its plants, but could produce 100,000 of a single model in each plant, using the same amount of capital and labour. This is a situation of increasing returns to scale, and in this instance the scale economies are what make gains from trade possible - as opposed to any innate comparative advantage between the economies. If trade barriers against the shipment of autos across national boundaries can be eliminated, then Hunda can take advantage of scale economies in each plant and increase its total production without using more capital and labour.

As this example implies, an opening up of trade increases the potential market size, and producers who experience increasing returns to scale stand to benefit from an enlarged market because their potential unit costs fall. Returns to scale are not limited to finished goods. Returns to scale characterize the production of many intermediate goods, which are goods used to produce other final goods or services. Manufacturers rarely produce all of the components entering their final products; they have **supply chains** for components that comprise numerous suppliers. In the automotive industry transmissions, gearboxes and seats are such **intermediate goods**. If either returns to scale, or comparative advantage, characterize their supply then there are gains to trade in these goods. In the context of the North American Free Trade Agreement (NAFTA). and its most recent form (the US Mexico and Canada Agreement - USMCA), automotive parts as well as automobiles can be shipped free of tariffs across borders provided that they satisfy **regional value content** (RVC) rules. We observe, for example, transmissions being produced in Ontario and seats in Mexico, and these goods are also shipped freely within North America provided they satisfy the RVC rules. Scale economies characterize transmission manufacture, and comparative advantage characterizes seat production – labour costs are lower in Mexico, and seats are labor intensive.

Content requirements apply to some goods under the NAFTA/USMCA. In the case of vehicles and their components, NAFTA required a 62.5% regional value content - that is, to be imported into one of the NAFTA signatories free of tariffs, 62.5% of the vehicle value had to be attributable to production in one of the three economies. This requirement was designed to prevent the members from using an excessive amount of components from low-wage economies and thereby undermine production of parts and vehicles in North America. The USMCA raises (by the year 2023) the regional value component (RVC) to 75% for most vehicles (70% for heavy trucks), and the RVC to between 65% and 75% on vehicle parts.

> **Supply chain**: denotes the numerous sources for intermediate goods used in producing a final product

> **Intermediate good**: one that is used in the production of final output

> **Regional value content**: requires that a specified percentage of the final value of a product originate in the economies covered in the Agreement.

Dynamic gains from trade

The term **dynamic gains** denotes the potential for domestic producers to increase productivity as a result of competing with, and learning from, foreign producers.

> **Dynamic gains**: the potential for domestic producers to increase productivity by competing with, and learning from, foreign producers.

Production processes in reality are seldom static. Innovation is constant in the modern world, and innovation is manifested in the form of productivity improvements. An economy's production possibility frontier is determined by its endowments of capital and labour and also the efficiency with which it uses those productive factors. **Total factor productivity** defines how efficiently the factors of production are combined. Research suggests that in developed economies this productivity increases by about 1% per annum. This means that more output can be produced using the same amounts of capital and labour because production is being carried out more efficiently. In graphical terms, such productivity improvements effectively push out an economy's production possibility frontier by 1% per annum. For economies in the process of development, this productivity growth may be as high as 3% or 4% per annum – for the reason that these economies can observe and learn from economies that are ahead of it technologically.

Freer trade forces domestic firms to compete with foreign firms that may be more productive. Domestic firms that can learn and adapt to competition by becoming more efficient will survive, firms that cannot adapt will not. Inevitably, there will be winners and losers in the production sector of the economy, whereas in the consumption sector most consumers should be winners.

> **Total factor productivity**: a measure of how efficiently the factors of production are combined.

15.5 Trade barriers: Tariffs, subsidies and quotas

Despite the many good arguments favoring free or relatively free trade, we observe numerous trade barriers. These barriers come in several forms. A **tariff** is a tax on an imported product that is designed to limit trade and generate tax revenue. It is a barrier to trade. An import **quota** is a limitation on imports; other **non-tariff barriers** take the form of product content requirements, and **subsidies**. By raising the domestic price of imports, a tariff helps domestic producers but hurts domestic consumers. Quotas and other non-tariff barriers have similar impacts.

> A **tariff** is a tax on an imported product that is designed to limit trade in addition to generating tax revenue.
>
> A **quota** is a quantitative limit on an imported product.
>
> A **trade subsidy** to a domestic manufacturer reduces the domestic cost and limits imports.
>
> **Non-tariff barriers**, such as product content requirements, limit the gains from trade.

Application Box 15.2: Tariffs – the national policy of J.A. MacDonald

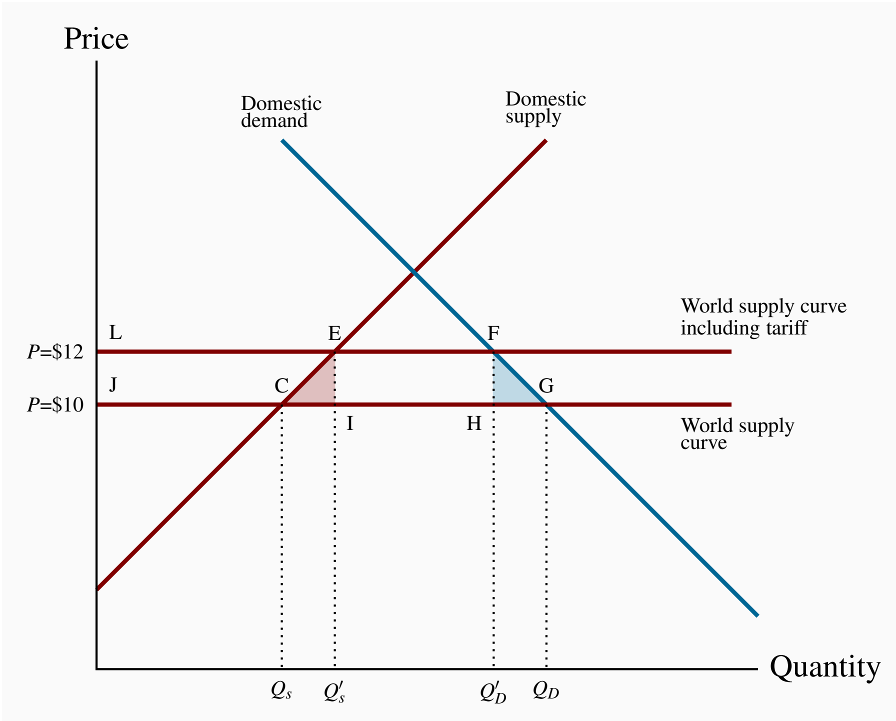
In Canada, tariffs were the main source of government revenues, both before and after Confederation in 1867 and up to World War I. They provided 'incidental protection' for domestic manufacturing. After the 1878 federal election, tariffs were an important part of the National Policy introduced by the government of Sir John A. MacDonald. The broad objective was to create a Canadian nation based on east-west trade and growth.

This National Policy had several dimensions. Initially, to support domestic manufacturing, it increased tariff protection on foreign manufactured goods, but lowered tariffs on raw materials and intermediate goods used in local manufacturing activity. The profitability of domestic manufacturing improved. But on a broader scale, tariff protection, railway promotion, Western settlement, harbour development, and transport subsidies to support the export of Canadian products were intended to support national economic development. Although reciprocity agreements with the United States removed duties on commodities for a time, tariff protection for manufactures was maintained until the GATT negotiations of the post-World War II era.

Tariffs

Figure 15.4 describes how tariffs operate. We can think of this as the Canadian wine market—a market that is heavily taxed in Canada. The world price of Cabernet Sauvignon is, let us say, \$10 per bottle, and this is shown by the horizontal world supply curve at that price. To maintain simplicity, we will neglect any taxes on alcohol here other than the tax represented by the tariff. The international supply curve is horizontal because the domestic market accounts for only a small part of the world demand for wine: we are sufficiently small that international producers can supply us with any amount we wish to buy at the world price. The Canadian demand for this wine is given by the demand curve D, and Canadian suppliers have a supply curve given by S (Canadian Cabernet is assumed to be of the same quality as the imported variety in this example). The effective supply curve in the Canadian market is now BCM. At a price of \$10, Canadian consumers wish to buy Q_D litres, and domestic producers wish to supply Q_S litres. The gap between domestic supply Q_S and domestic demand Q_D is filled by imports. This is the *free trade equilibrium*.

Figure 15.4: Tariffs and trade



At a world price of \$10 the domestic quantity demanded is Q_D. Of this amount Q_s is supplied by domestic producers and the remainder by foreign producers. A tariff increases the world price to \$12. This reduces demand to Q_D'; the domestic component of supply increases to Q_s'. Of the total loss in consumer surplus (LFGJ), tariff revenue equals EFHI, increased surplus for domestic suppliers equals LECJ, and the deadweight loss is therefore the sum of the triangular areas CEI and HFG.

If the government now imposes a 20 percent tariff on imported wines (or a \$2 per bottle tax), foreign wine sells for \$12 a bottle, inclusive of the tariff. The effective supply curve in the Canadian market becomes BEK. The tariff raises the domestic 'tariff-inclusive' price above the world price, and this shifts the international supply curve of this wine upwards. By raising wine prices in the domestic market, the tariff protects domestic producers by raising the domestic price at which imports become competitive. Those domestic suppliers who were previously not quite competitive at a global price of \$10 are now competitive. The total quantity demanded falls from Q_D to Q_D' at the new equilibrium F. Domestic producers supply the amount Q_S' and imports fall to the amount $(Q_D' - Q_S')$. Reduced imports are partly displaced by those domestic producers who can supply at prices between \$10 and \$12. Hence, imports fall both because total consumption falls and because domestic suppliers can displace some imports under the protective tariff; the amount $(Q_S' - Q_S)$.

Since the tariff is a type of tax, its impact in the market depends upon the elasticities of supply and demand, (as illustrated in Chapters 4 and 5). The more elastic is the demand curve, the more a given tariff reduces imports. In contrast, if it is inelastic the quantity of imports declines less.

Costs and benefits of a tariff

The costs of a tariff come from the higher price to consumers, but this is partly offset by the tariff revenue that goes to the government. This tariff revenue is a benefit and can be redistributed to consumers or spent on goods from which consumers derive a benefit. But there are also efficiency costs associated with tariffs—deadweight losses, as we call them. These are the real costs of the tariff, and they arise because the marginal cost of production does not equal the marginal benefit to the consumer. Let us see how these concepts apply with the help of Figure 15.4.

Consumer surplus is the area under the demand curve and above the equilibrium market price. It represents the total amount consumers would have been willing to pay for the product, but did not have to pay, at the equilibrium price. It is a measure of consumer welfare. The tariff raises the market price and reduces this consumer surplus by the amount $LFGJ$. This area measures by how much domestic consumers are worse off as a result of the price increase caused by the tariff. But this is not the net loss for the whole domestic economy, because the government obtains some tax revenue and domestic producers get more revenue and profit.

Government revenue accrues from the domestic sales of imports. On imports of $(Q_D' - Q_S')$, tax revenue is $EFHI$. Then, domestic producers obtain an additional profit of $LECJ$—the excess of additional revenue over their cost per additional bottle. If we are not concerned about who gains and who loses, then there is a net loss to the domestic economy equal to the areas CEI and HFG.

The area HFG is the consumer side measure of deadweight loss. At the quantity Q_D', the production cost of an additional bottle is less than the value placed on it by consumers; and, by not having those additional bottles supplied, consumers forgo a potential gain. The area CEI tells us that when supply by domestic higher-cost producers is increased, and supply of lower-cost foreign producers is reduced, the corresponding resources are not being used efficiently. The sum of the areas CEI and HFG is therefore the total deadweight loss of the tariff.

In the real world we should also be interested in the magnitude of the financial amounts involved here: In particular, how much more do consumers pay with the tariff in place, relative to the additional amounts going to domestic suppliers/corporations? How much tax revenue is generated? How many jobs are created domestically as a result of 'distorting' the market? Regardless of the magnitude of the two deadweight loss areas, which represent the net cost of the tariff, we should be interested in whether the owners of capital gain at the expense of consumers.

Tariffs by country of origin - trade diversion

The imposition of tariffs is governed by the World Trade Organization (WTO). Tariffs are permitted under the WTO rules in specific circumstances: if a particular economy is deemed to be subsidizing exports, and those exports have employment impacts on the destination economy, then a 'retaliatory' tariff may be imposed. A related justification is **dumping**. It is frequently difficult to prove subsidization or dumping by an exporting economy. An example of such a tariff was one placed by the US on washing machines originating in China in 2016, on the basis of a dumping claim by the United States. The immediate result of this was that the manufacturers located in

China switched most of their production to other plants they owned in Vietnam and Thailand. In this particular instance there was virtually no impact on the retail price of washing machines in the US.

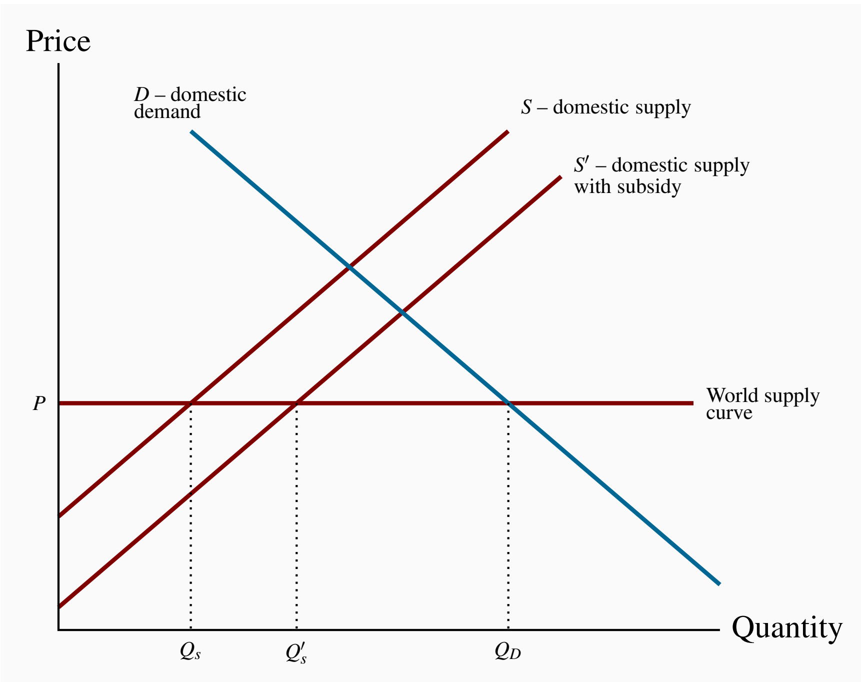
> **Dumping** is a predatory practice, based on artificially low costs aimed at driving out domestic producers.

The traditional theory of tariffs described in 15.4 implicitly assumes that production and employment increase in the importing economy as a result of domestic production displacing imported goods. This analysis assumes that the tariff is imposed on a particular commodity, regardless of its economy of origin.

Production subsidies

Figure 15.5 illustrates the effect of a subsidy to a domestic supplier. As in Figure 15.4, the amount Q_D is demanded in the free trade equilibrium and, of this, Q_S is supplied domestically. With a subsidy per unit of output sold, the government can reduce the supply cost of the domestic supplier, thereby shifting the supply curve downward from S to S'. In this illustration, the total quantity demanded remains at Q_D, but the domestic share increases to Q'_S.

Figure 15.5: Subsidies and trade



With a world supply price of P, a domestic supply curve S, and a domestic demand D, the amount Q_D is purchased. Of this, Q_s is supplied domestically and $(Q_D - Q_s)$ by foreign suppliers. A per-unit subsidy to domestic suppliers shifts their supply curve to S', and increases their market share to Q'_s.

The new equilibrium represents a misallocation of resources. When domestic output increases from Q_S to Q'_S, a low-cost international producer is being replaced by a higher cost domestic supplier; the domestic supply curve *S* lies above the international supply curve *P* in this range of output.

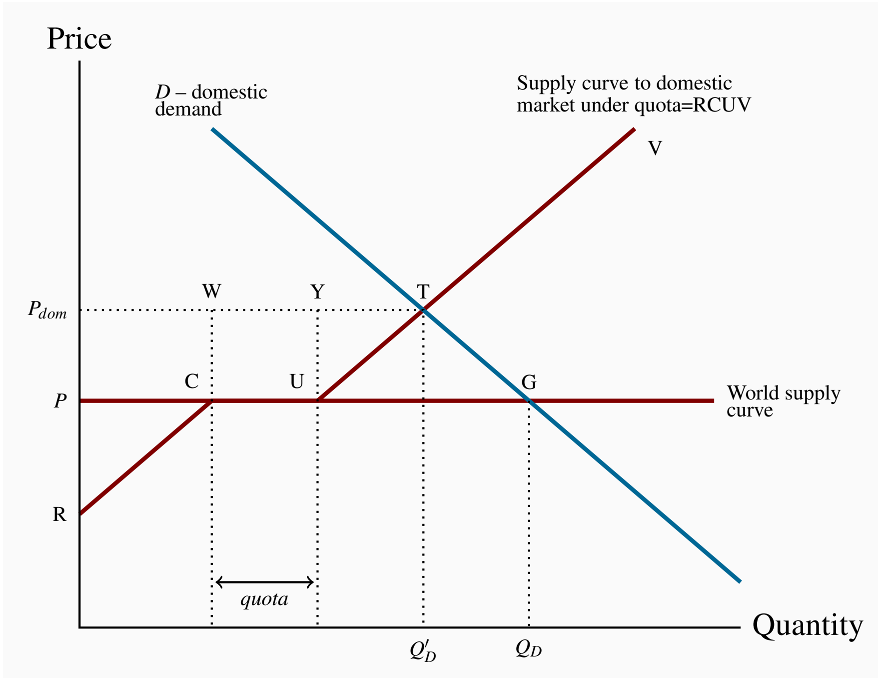
Note that this example deals with a subsidy to domestic suppliers who are selling in the domestic market. It is not a subsidy to domestic producers who are selling in the international market – an export subsidy.

This subsidy comes with a cost to the domestic economy: Taxpayers-at-large must pay higher taxes to support this policy; and each dollar raised in tax itself has a deadweight loss, as we examined in Chapter 5.

Quotas

A quota is a limit placed upon the amount of a good that can be imported. Consider Figure 15.6, where again there is a domestic supply curve coupled with a world price of *P*. Rather than imposing a tariff, the government imposes a quota that restricts imports to a physical amount denoted by the distance *quota* on the quantity axis. The supply curve facing domestic consumers then has several segments to it. First it has the segment *RC*, reflecting the fact that domestic suppliers are competitive with world suppliers up to the amount *C*. Beyond this output, world suppliers can supply at a price of *P*, whereas domestic suppliers cannot compete at this price. Therefore the supply curve becomes horizontal, but only *up to the amount permitted under the quota*—the quantity *CU* corresponding to *quota*. Beyond this amount, international supply is not permitted and therefore additional amounts are supplied by the (higher cost) domestic suppliers. Hence the supply curve to domestic buyers becomes the supply curve from the domestic suppliers once again.

Figure 15.6: Quotas and trade



At the world price P, plus a *quota*, the supply curve becomes RCUV. This has three segments: (i) domestic suppliers who can supply below P; (ii) *quota*; and (iii) domestic suppliers who can only supply at a price above P. The quota equilibrium is at T, with price P_{dom} and quantity Q'_D; the free-trade equilibrium is at G. Of the amount Q'_D, *quota* is supplied by foreign suppliers and the remainder by domestic suppliers. The quota increases the price in the domestic market.

The resulting supply curve yields an equilibrium quantity Q'_D. There are several features to note about this equilibrium. First, the quota pushes the domestic price above the world price (P_{dom} is greater than P) because low-cost international suppliers are partially supplanted by higher-cost domestic suppliers. Second, if the quota is chosen 'appropriately', the same domestic market price could exist under the quota as under the tariff in Figure 15.4. Third, in contrast to the tariff case, the government obtains no tax revenue from the quotas. The higher market price under a quota means that the price per unit received by foreign suppliers is now P_{dom} rather than P. *De facto*, instead of tax revenue being generated in the importing economy, the foreign supplier benefits from a higher price. Fourth, inefficiencies are associated with the equilibrium at Q'_D. These inefficiencies arise because the lower-cost international suppliers are not permitted to supply the amount they would be willing to supply at the quota-induced market equilibrium. In other words, more efficient producers are being squeezed out of the market by quotas that make space for less-efficient producers.

Application Box 15.3: Cheese quota in Canada

In 1978 the federal government set a cheese import quota for Canada at just over 20,000 tonnes. This quota was implemented initially to protect the interests of domestic suppliers. Despite a strong growth in population and income in the intervening decades, the import quota has remained unchanged. The result is a price for cheese that is considerably higher than it would otherwise be. The quotas are owned by individuals and companies who have the right to import cheese. The quotas are also traded among importers, at a price. Importers wishing to import cheese beyond their available quota pay a tariff of about 250 percent. So, while the consumer is the undoubted loser in this game, who gains?

First the suppliers gain, as illustrated in Figure 15.6. Canadian consumers are required to pay high-cost domestic producers who displace lower-cost producers from overseas. Second, the holders of the quotas gain. With the increase in demand for cheese that comes with higher incomes, the domestic cheese price increases over time and this in turn makes an individual quota more valuable.

In the 2018 United States Mexico Canada Agreement, a slight increase in access to Canadian markets was granted in return for a corresponding increase in access to the US market.

15.6 The politics of protection

Objections to imports are frequent and come from many different sectors of the economy. In the face of the gains from trade which we have illustrated in this chapter, why do we observe such strong opposition to imported goods and services?

Structural change and technology

In a nutshell the answer is that, while consumers in the aggregate gain from the reduction of trade barriers, and there is a net gain to the economy at large, some *individual sectors of the economy lose out*. Not surprisingly the sectors that will be adversely affected are vociferous in lodging their objections. Sectors of the economy that cannot compete with overseas suppliers generally see a reduction in jobs. This has been the case in the manufacturing sector of the Canadian and US economies in recent decades, as manufacturing and assembly has flown off-shore to Asia and Mexico where labour costs are lower. Domestic job losses are painful, and frequently workers who have spent decades in a particular job find reemployment difficult, and rarely get as high a wage as in their displaced job.

Such job losses are reflected in calls for tariffs on imports from China, for example, in order to 'level the playing field' – that is, to counter the impact of lower wages in China. Of course it is precisely because of lower labour costs in China that the Canadian consumer benefits.

In Canada we deal with such dislocation first by providing unemployment payments to workers, and by furnishing retraining allowances, both coming from Canada's Employment Insurance program. While such support does not guarantee an equally good alternative job, structural changes in the economy, due to both internal and external developments, must be confronted. For example, the information technology revolution made tens of thousands of 'data entry' workers redundant. Should producers have shunned the technological developments which increased their productivity dramatically? If they did, would they be able to compete in world markets?

While job losses feature heavily in protests against technological development and freer trade, most modern economies continue to grow and create more jobs in the service sector than are lost in the manufacturing sector. Developed economies now have many more workers in service than manufacture. Service jobs are not just composed of low-wage jobs in fast food establishments – 'Mcjobs', they are high paying jobs in the health, education, legal, financial and communications sectors of the economy.

Successful lobbying and concentration

While efforts to protect manufacture have not resulted in significant barriers to imports of manufactures, objections in some specific sectors of the economy seem to be effective worldwide. One sector that stands out is agriculture, where political conditions are conducive to the continuance of protection and what is called 'supply management' – domestic production quotas. The reason for 'successful' supply limitation appears to rest in the geographic concentration of potential beneficiaries of such protection and the scattered beneficiaries of freer trade on the one hand, and the costs and benefits of political organization on the other: Farmers tend to be concentrated in a limited number of rural electoral ridings and hence they can collectively have a major impact on electoral outcomes. Second, the benefits that accrue to trade restriction are heavily concentrated in the economy – keep in mind that about two percent of the population lives on farms, or relies on farming for its income. By contrast the costs on a per person scale are small, and are spread over the whole population. Thus, in terms of the costs of political organization, the incentives for consumers are small, but the incentives for producers are high.

In addition to the differing patterns of costs and benefits, rural communities tend to be more successful in pushing trade restrictions based on a 'way-of-life' argument. By permitting imports that might displace local supply, lobbyists are frequently successful in convincing politicians that long-standing way-of-life traditions would be endangered, even if such 'traditions' are accompanied by monopolies and exceptionally high tariffs.

Valid trade barriers: Infant industries and dumping?

An argument that carries both intellectual and emotional appeal to voters is the 'infant industry' argument. It goes as follows: New ventures and sectors of the economy may require time before that can compete internationally. Scale economies may be involved, for example, and time may be required for producers to expand their scale of operation, at which time costs will have fallen to international (i.e. competitive) levels. In addition, learning-by-doing may be critical in more high-tech sectors and, once again, with the passage of time costs should decline for this reason also.

The problem with this stance is that these ‘infants’ have insufficient incentive to ‘grow up’ and become competitive. A protection measure that is initially intended to be temporary can become permanent because of the potential job losses associated with a cessation of the protection to an industry that fails to become internationally competitive. Furthermore, employees and managers in protected sectors have insufficient incentive to make their production competitive if they realize that their government will always be there to protect them.

In contrast to the infant industry argument, economists are more favourable to restrictions that are aimed at preventing dumping.

Dumping may occur either because foreign suppliers choose to sell at artificially low prices (prices below their break-even price for example), or because of surpluses in foreign markets resulting from oversupply. For example, if, as a result of price support in its own market, a foreign government induced oversupply in butter and it chose to sell such butter on world markets at a price well below the going (‘competitive’) world supply price, such a sale would constitute dumping. Alternatively, an established foreign supplier might choose to enter our domestic market by selling its products at artificially low prices, with a view to driving domestic competition out of the domestic market. Having driven out the domestic competition it would then be in a position to raise prices. This is predatory pricing as explored in the last chapter. Such behaviour differs from a permanently lower price on the part of foreign suppliers. This latter may be welcomed as a gain from trade, whereas the former may generate no gains and serve only to displace domestic labour and capital.

Protectionism in the age of pandemics

The year 2020 will be remembered in history as the year of the coronavirus pandemic. An uncountable number of men and women died all across the globe as a result of contracting COVID-19, the respiratory disorder brought on by an attack of the coronavirus. In the absence of a vaccine, health authorities the world over implemented a twin policy of social distancing and quarantining (or self-isolation). The world economy went into a tailspin, as huge fractions of the labor force were laid off. Trade patterns were disrupted and serious shortages of personal protection equipment (PPE - masks, visors, gowns), ventilators and drugs emerged. The world demand for PPE and ventilators skyrocketed. But the production of PPE was concentrated in China; most western economies did not have the necessary productive capacity to supply even non-pandemic requirements. Bidding wars erupted amongst countries and hospitals as they vied for supply, while domestic producers of some products added to their production capacity.

Following this chaos, we ask if self-sufficiency would not be a better model than open trade. Would a world where each country ensured it had the production capacity to produce these necessities in times of emergency not be superior to one where global supply chains characterize everything from computers to generic drugs? India is a major producer of generic drugs and the components for such drugs. The demand for anti-biotics and pain killers also rocketed upwards with the pandemic.

There is more than one way to plan for a pandemic, and such planning should not involve a generalized move to self-sufficiency on the part of the global economy. One strategy is to build up inventories of PPE and ventilators domestically. This is costly, but for the most part feasible. It does not represent a complete solution because technology changes will make 30-year old ventilators sitting in inventory redundant for the next pandemic. In addition, most medications have a limited shelf life. Hence one solution is to maintain and rotate substantial inventories of emergency equipment using existing supply chains, and benefit from the efficiencies that are built into these chains.

A second option is to maintain excess production capacity on the part of domestic manufacturers of critical pandemic products. Maintaining such capacity should be considered at least partially as a social cost; pandemics ravage societies, not just individuals, and therefore society should undertake part of the cost of insuring against them.

A more general argument against global trade comes in the form of protecting food supplies. In the early 2000s an increase in global cereal prices led some economies to limit exports of specific crops on account of the fact that global demand was pushing prices to a level that low-income consumers could not afford. But such a policy may threaten consumers in other low-income economies whose demands have not changed in a context of reduced supplies. The reality is that world food supply is adequate for world consumption, even in the presence of disruptions. It is also the case that certain economies have huge advantages in producing specific kinds of food. For example, Canada, the US and the Ukraine produce cereals very economically. Mountainous regions are unsuitable for this production. It would benefit no economy for these economies to lower their production of grains to the point where they produced only enough for their own production. By the same reasoning, warmer climates produce fruits, coffee beans, olives etc that cannot easily be produced in many regions suited to wheat. The gains to specialization in the world economy are enormous. Where food shortages occur we frequently encounter the scourges of drought or war or political upheaval, and these conditions inhibit the distribution of foodstuffs.

What about supply chains? If motherboards produced in China are not being exported in sufficient quantities then indeed production of computers in North America will suffer. But to infer from this that North America should decide to produce all of its computer components in North America is illogical. First, in the time of a pandemic, if certain economies in the supply chain are on lockdown, we cannot be sure that the domestic economy would not be on lockdown simultaneously. Second, the cost to moving the production of all computer parts to North America would likely double the cost of computer hardware - including cell-phones. Perhaps a disruption to our supply chains is something we need to bear in extraordinary times. In case it requires emphasis, most producers in supply chains have incentives to produce and sell. If they do not they will die economically.

The energy sector of every economy is impacted with the outbreak of a pandemic. This is because the demand for fuel (primarily oil) declines following policies of social distancing, limits on permissible travel, and the closure of some production facilities that depend upon oil. In North America, as we saw in Chapter 4 earlier, the price of oil declined from US $60 per barrel to US $20 in the space of two months in early 2020. Since production costs are higher in both Canada and much of the US than in Saudi Arabia, the North Sea and Russia, producers in North America were squeezed. Many were no longer able to cover their full production costs, and forced to cease drilling and recovering oil. Inevitably there was a clamor for protection. Producers sought tariffs on competing oil: Tariffs would increase the price of cheaper-to-produce foreign oil and enable domestic producers to survive.

While protection might seem like a 'sensible' policy in this instance, the fact is that unilateral tariffs usually invite reprisals, and raise the danger of a trade war with ever-expanding counter protectionism. In contrast to the case of a *shortage* of medical supplies, the energy sector in Canada suffered from a *glut* of world oil supply. The domestic issue is not about the health of consumers (as in the case of medical supplies), it is about the health of producers.

To conclude: a pandemic is a profoundly serious event and such events inflict major costs on all societies. There are no magic bullets in the form of low-cost ideal economic policies to counter viral warfare. The key to policy making is to recognize constraints and recognize an attack as soon as possible. A wholesale move to insulate the domestic economy is ill-conceived. Comparative advantage confers enormous benefits to all nations. Specific policies should take the form of inventory management and excess production capacity in specific sectors of the economy.

15.7 Institutions governing trade

In the nineteenth century, world trade grew rapidly, in part because the leading trading nation at the time—the United Kingdom—pursued a vigorous policy of free trade. In contrast, US tariffs averaged about 50 percent, although they had fallen to around 30 percent by the early 1920s. As the industrial economies went into the Great Depression of the late 1920s and 1930s, there was pressure to protect domestic jobs by keeping out imports. Tariffs in the United States returned to around 50 percent, and the United Kingdom abandoned the policy of free trade that had been pursued for nearly a century. The combination of world recession and increasing tariffs led to a disastrous slump in the volume of world trade, further exacerbated by World War II.

The WTO and GATT

After World War II, there was a collective determination to see world trade restored. Bodies such as the International Monetary Fund and the World Bank were set up, and many countries signed the General Agreement on Tariffs and Trade (GATT), a commitment to reduce tariffs progressively and dismantle trade restrictions.

Under successive rounds of GATT, tariffs fell steadily. By 1960, United States tariffs were only one-fifth of their level at the outbreak of the War. In the United Kingdom, the system of wartime quotas on imports had been dismantled by the mid-1950s, after which tariffs were reduced by nearly half in the ensuing 25 years. Europe as a whole moved toward an enlarged European Union in which tariffs between member countries have been abolished. By the late 1980s, Canada's tariffs had been reduced to about one-quarter of their immediate post-World War II level.

The GATT Secretariat, now called the World Trade Organization (WTO), aims both to dismantle existing protection that reduces efficiency and to extend trade liberalization to more and more countries. Tariff levels throughout the world are now as low as they have ever been, and trade liberalization has been an engine of growth for many economies. The consequence has been a substantial growth in world trade.

NAFTA, the USMCA, the EU, the CETA, and the TPP

In North America, policy since the 1980s has led to a free trade area that covers the flow of trade between Canada, the United States, and Mexico. The Canada/United States free trade agreement (FTA) of 1989 expanded in 1994 to include Mexico in the North American Free Trade Agreement (NAFTA). The objective in both cases was to institute freer trade between these countries in most goods and services. This meant the elimination or reduction of tariffs and non-tariff barriers over a period of years, with a few exceptions in specific products and cultural industries. A critical component of the Agreement was the establishment of a dispute-resolution mechanism, under which disputes would be resolved by a panel of 'judges' nominated from the member economies. Evidence of the success of these agreements is reflected in the fact that Canadian exports have grown to more than 30 percent of GDP, and trade with the United States accounts for the lion's share of Canadian trade flows. NAFTA was updated and replaced in 2018 and the new agreement is termed the United States Mexico Canada Agreement.

The European Union was formed after World War II, with the prime objective of bringing about a greater degree of *political* integration in Europe. Two world wars had laid waste to their economies and social fabric. Closer economic ties and greater trade were seen as the means of achieving this integration. The Union was called the "Common Market" for much of its existence. The Union originally had six member states, and as of 2019 the number is 28, with several other candidate countries in the process of application, most notably Turkey. The European Union (EU) has a secretariat and parliament in Bruxelles. The UK intends to exit the EU as of late 2019.

Canada has concluded a free trade agreement with the European Union that is termed the Comprehensive Economic and Trade Agreement (CETA). It has the objective of implementing free trade between the two negotiating parties, though there remain some exceptions, for example agriculture.

The Comprehensive and Progressive Agreement for Trans-Pacific Partnership (CPTPP) is a trading agreement between Canada and ten other Pacific-Rim economies that came into being in 2018. Negotiations for a Trans Pacific Partnership treaty were complete by 2016. Those negotiations involved 12 Pacific Rim economies including Canada and the United States, but excluding China. The Obama presidency appeared ready to sign the treaty, however the Trump presidency (and also the Democratic candidate for president of the US, Hillary Clinton) decided that the Partnership was not in the interests of the United States and withdrew its affiliation. The remaining 11 economies reached an agreement to implement the partnership in December 2018.

Key Terms

Autarky denotes the no-trade situation.

Principle of comparative advantage states that even if one country has an absolute advantage in producing both goods, gains to specialization and trade still materialize, provided the opportunity cost of producing the goods differs between economies.

Terms of trade define the rate at which goods trade internationally.

Consumption possibility frontier defines what an economy can consume after production specialization and trade.

Intra-industry trade is two-way international trade in products produced within the same industry.

Intra-firm trade is two-way trade in international products produced within the same firm.

Supply chain: denotes the numerous sources for intermediate goods used in producing a final product.

Intermediate good: one that is used in the production of final output.

Content requirement: requires that a specified percentage of the final value of a product originate in the producing economy.

Dynamic gains: the potential for domestic producers to increase productivity by competing with, and learning from, foreign producers.

Total factor productivity: how efficiently the factors of production are combined.

Tariff is a tax on an imported product that is designed to limit trade in addition to generating tax revenue. It is a barrier to trade.

Quota is a quantitative limit on an imported product.

Trade subsidy to a domestic manufacturer reduces the domestic cost and limits imports.

Non-tariff barriers, such as product content requirements, limits the gains from trade.

Dumping is a predatory practice, based on artificial costs aimed at driving out domestic producers.

EXERCISES FOR CHAPTER 15

Exercise 15.1 The following table shows the labour input requirements to produce a bushel of wheat and a litre of wine in two countries, Northland and Southland, on the assumption of constant cost production technology – meaning that the production possibility curves in each are straight lines. You can answer this question either by analyzing the table or developing a graph similar to Figure 15.1, assuming each economy has 4 units of labour.

Labour requirements per unit produced		
	Northland	Southland
Per bushel of wheat	1	3
Per litre of wine	2	4

(a) Which country has an absolute advantage in the production of both wheat and wine?

(b) What is the opportunity cost of wheat in each economy? Of wine?

(c) What is the pattern of comparative advantage here?

(d) Suppose the country with a comparative advantage in wine reduces wheat production by one bushel and reallocates the labour involved to wine production. How much additional wine does it produce?

Exercise 15.2 Canada and the United States can produce two goods, xylophones and yogurt. Each good can be produced with labour alone. Canada requires 60 hours to produce a ton of yogurt and 6 hours to produce a xylophone. The United States requires 40 hours to produce the ton of yogurt and 5 hours to produce a xylophone.

(a) Describe the state of absolute advantage between these economies in producing goods.

(b) In which good does Canada have a comparative advantage? Does this mean the United States has a comparative advantage in the other good?

(c) Draw the production possibility frontier for each economy to scale on a diagram, assuming that each economy has an endowment of 240 hours of labour, and that the PPFs are linear.

(d) On the same diagram, draw Canada's consumption possibility frontier on the assumption that it can trade with the United States at the United States' rate of transformation.

(e) Draw the US consumption possibility frontier under the assumption that it can trade at Canada's rate of transformation.

Exercise 15.3 The domestic demand for bicycles is given by $P = 36 - 0.3Q$. The foreign supply is given by $P = 18$ and domestic supply by $P = 16 + 0.4Q$.

(a) Illustrate the market equilibrium on a diagram, and illustrate the amounts supplied by domestic and foreign suppliers in equilibrium.

(b) If the government now imposes a tariff of \$6 per unit on the foreign good, illustrate the impact geometrically.

(c) In the diagram, illustrate the area representing tariff revenue.

(d) *Optional*: Compute the price and quantity in equilibrium with free trade, and again in the presence of the tariff.

Exercise 15.4

(a) In Exercise 15.3, illustrate graphically the deadweight losses associated with the imposition of the tariff.

(b) Illustrate on your diagram the additional amount of profit made by the domestic producer as a result of the tariff. [*Hint*: Refer to Figure 15.4 in the text.]

Exercise 15.5 The domestic demand for office printers is given by $P = 40 - 0.2Q$. The supply of domestic producers is given by $P = 12 + 0.1Q$, and international supply by $P = 20$.

(a) Illustrate this market geometrically.

(b) If the government gives a production subsidy of \$2 per unit to domestic suppliers in order to increase their competitiveness, illustrate the impact of this on the domestic supply curve.

(c) Illustrate geometrically the cost to the government of this scheme.

Exercise 15.6 Consider the data underlying Figure 15.1. Suppose, from the initial state of comparative advantage, where Canada specializes in fish and the US in vegetable, we have a technological change in fishing. The US invents the multi-hook fishing line, and as a result can now produce 64 units of fish with the same amount of labour, rather than the 40 units it could produce before the technological change. This technology does not spread to Canada however.

(a) Illustrate the new PPF for the US in addition to the PPF for Canada.

(b) What is the new opportunity cost (number of fish) associated with one unit of V?

(c) Has comparative advantage changed here – which economy should specialize in the production of each good?

Exercise 15.7 The following are hypothetical (straight line) production possibilities tables for Canada and the United States. For each line required, plot any two or more points on the line.

Canada					United States				
	A	B	C	D		A	B	C	D
Peaches	0	5	10	15	**Peaches**	0	10	20	30
Apples	30	20	10	0	**Apples**	15	10	5	0

(a) Plot Canada's production possibilities curve.

(b) Plot the United States' production possibilities curve.

(c) What is each country's cost ratio of producing peaches and apples?

(d) Which economy should specialize in which product?

(e) Plot the United States' trading possibilities curve (by plotting at least 2 points on the curve) if the actual terms of the trade are 1 apple for 1 peach.

(f) Plot the Canada' trading possibilities curve (by plotting at least 2 points on the curve) if the actual terms of the trade are 1 apple for 1 peach.

(g) Suppose that the optimum product mixes before specialization and trade were B in the United States and C in Canada. What are the gains from specialization and trade?

Part Seven

Introduction to Macroeconomics

The four chapters in this part of the text introduce and develop the expenditure side of a basic traditional macroeconomic model. Chapter 16 explains the measurement of macroeconomic activity and performance. Then Chapter 17 introduces an aggregate demand and supply model of national output, the general price level and business cycles. Market based aggregate expenditure components that determine aggregate demand, when prices, interest rates and exchange rates are constant, are modelled in Chapter 18. Chapter 19 extends the expenditure model to include Government expenditure, taxes, budgets, public debt and basic fiscal policy.

Chapter 16

Economic activity & performance

In this chapter we will explore:

Macroeconomics is focused on three key indicators of the economy's performance and the underlying explanations for their behaviour. The indicators are:

1. The rate of growth of real national income.
2. The rate of inflation.
3. The rate of unemployment.

Other aspects of the economy like interest rates, foreign exchange rates, wage rates, government budgets, capital investment, commodity prices, housing and so forth are important to macroeconomic analysis because they work to determine performance as measured by these three indicators.

There are, however, longer term perspectives, decades and beyond, that involve important but different economic issues. The sustained growth of real income per person is of major interest. But underlying that growth are questions raised by: climate change and energy sources; demographic trends that change labour force participation, dependency ratios, social policies, and immigration policy; changes in the industrial structure of employment and incomes; changing national patterns of economic activity and international trade policy and patterns; and that list is not exhaustive.

These are all important matters but they lie outside the scope of an introduction to the principles of macroeconomics. We will focus on explaining short term economic observations and theory that determine the three key performance indicators listed above, which in turn have important impacts on standards of living and underlie longer term issues.

Macroeconomics involves complex linkage and feedback effects that tie economic conditions and economic policy to economic performance. Macroeconomic theories and models attempt to capture this complexity. They seek to understand and explain the causes of short run changes in economic performance and the role for economic policy.

For example, the persistent effects of what seemed to be a short term local crisis in the US housing and financial markets triggered the Great Recession of 2009. International linkages among financial markets spread the effects across European and other financial markets. Government bailouts of major banks and monetary and fiscal stimulus used to fight falling output and employment resulted in unprecedented government deficits and historically low interest rates. International financial and fiscal linkages were much stronger than expected initially.

In Canada, the interest rates set by the Bank of Canada were reduced to the lowest historical level. Governments focused their budget policies on reducing or eliminating budget deficits caused by earlier economic conditions and policy decisions, despite pressing needs for infrastructure investment. The later collapse in crude oil and other commodity prices raised new concerns for domestic economic growth and employment.

But by mid-2016 energy and commodity prices had recovered and the new government's expansionary fiscal policy supported economic recovery. Growth in Canadian economic activity and employment increased steadily. Even so, by Bank of Canada estimates, the Canadian economy still remained below 'full employment' in the third quarter of 2019.

Government policy to contain COVID-19 will change all previous forecasts. Early estimates by the *Parliament Budget Office, April 2020* are real GDP growth of -5.1 percent and an unemployment rate of 12.4 percent by the fourth quarter of 2020. (See: Revised PBO Report RP 2020-004-S, Table 21.)

Macroeconomic theory and models emerged from an earlier major financial collapse and crisis followed by the depression years of the 1930s. Although today's economies are larger and more complex they still behave by the same basic principles.

To understand the different dimensions of economic activity, economic conditions and macroeconomic policies, we need a framework that captures how they are related and how they interact. Macroeconomics provides that framework, using consistent and comprehensive system of definitions for the measurement of economic activity provided by the national accounts.

16.1 Indicators of macroeconomic activity and performance

Output, price, and employment are three key dimensions of macroeconomic activity. Output is a measure of the total quantity of goods and services produced in the economy. It is also a measure of the incomes generated by that production. Price or the price level in macroeconomics is the weighted average of the market prices of all final goods and services produced. The price level reflects the costs of production in the economy. Employment is a measure of the number of jobs involved in the production of goods and services, or, in more refined terms, the number of hours of labour input required to produce the economy's output. Economic performance is judged by how these measures change over time.

Output and its rate of growth are measured in terms of **real gross domestic product (real GDP)**. It is the quantity of final goods and services produced in the economy in a specific time period, such as one year, measured in the market prices of a base year, 2012 for example. (It may also be called GDP in constant 2012 dollars). The production of goods and services generates incomes equal to the value of those goods and services. As a result, real GDP is also the real income in the economy and the quantity of goods and services the economy can afford to buy.

> **Real GDP**: the quantity of final goods and services produced by the economy in a specified time period.

In an economy with a growing population and labour force, growth in real GDP is necessary to maintain standards of living. In the Canadian economy, real GDP changes from year to year. By measuring real GDP in the prices of a base year, the changes seen in real GDP are the result of changes in the quantities of goods and services produced, and not the result of changes in prices. This distinction is important: Increased quantities of goods and services provide for increased standards of living in the economy, increases in prices do not. As a result, **economic growth** is defined as an increase in real GDP, and the annual **rate of economic growth** is the annual percentage change in real GDP. This is the first key indicator of economic performance.

The annual rate of growth in real GDP is calculated as follows:

$$\text{Rate of growth of real GDP} = \frac{\text{Real GDP}_{\text{year 2}} - \text{Real GDP}_{\text{year 1}}}{\text{Real GDP}_{\text{year 1}}} \times 100 \tag{16.1}$$

> **Economic growth**: an increase in real GDP.
>
> **Rate of economic growth**: the annual percentage change in real GDP.

Recent measures of real GDP in Canada provide an example of economic growth and the calculation of the rate of economic growth. In the fourth quarter of 2018, real GDP in Canada measured at annual rates in 2012 dollars was $2,079 billion. One year earlier, in the fourth quarter of 2017, real GDP in 2012 dollars was $2,037 billion. Using these data:

$$\text{Rate of growth of real GDP in 2018} = \frac{\$2,079 - \$2,037}{\$2,037} \times 100 = 2.06\%$$

The **price level** in the economy is a measure of the weighted average of prices of a wide variety of goods and services. The **Consumer Price Index (CPI)**, for example, compares the cost of a fixed basket of goods and services bought by the typical household at a specific time with the cost of that same basket of goods and services in the base year. It is the most widely used indicator of prices in Canada and is often referred to as the "cost of living."

> **Price level**: a measure of the average prices of all goods and services produced in the economy.
>
> **Price index**: a measure of the price level in one year compared with prices in a base year.
>
> **Consumer Price Index (CPI)**: a measure of the cost of living in any one year compared to the cost of living in a base year.

The Consumer Price Index is a more comprehensive measure of the change in prices from one year to the next, but the simple example in Example Box 16.1 illustrates the how such an index is constructed and what it tells us.

Example Box 16.1: Constructing a price index

University student weekly expenditure basket (Base year 2015)

	Quantity	2015 Price	2015 Cost	2019 Price	2019 Cost
Pizza	5	\$7.50	\$37.50	\$8.50	\$42.50
Hamburger	5	\$2.50	\$12.50	\$2.25	\$11.25
Coffee	10	\$1.00	\$10.00	\$1.25	\$12.50
Movies	1	\$10.00	\$10.00	\$8.00	\$8.00
Bus fare	7	\$1.50	\$10.50	\$1.85	\$12.95
Total			\$80.50		\$87.20

Suppose a survey of expenditures by university students in the year 2015 gives the information reported in the first three columns of the table. The cost of weekly expenditures on a basket of five items and the weight of each item in the total expenditure was \$80.50. Using 2015 as our base year, the cost of the basket in 2015 prices, \$80.50, has an index value of 100 $[(\$80.50/80.50) \times 100]$. In other words we have a Student Price Index:

$$\text{SPI}_{2015} = 100.0$$

Now in the last two columns of the table this same basket of goods and services in *the prices of 2019* would cost \$87.20. Then our SPI in 2019 would be:

$$(\text{Cost of basket in 2019})/(\text{Cost of basket in 2015}) \times 100 = (\$87.20)/(\$80.50) \times 100 = 108.3$$

The index tells us that even though the prices of some things went up and others went down the Student Price Index increased by 8.3%. This was the weighted average increase in prices and the increase in the cost of student expenditures.

Today, the base year for the Canadian consumer price index is 2002 with a value of 100. Statistics Canada uses a fixed basket classified under eight consumer expenditure categories. The weight or importance of each category is its share of expenditure as determined by consumer expenditure surveys. By visiting the Statistics Canada website, `https://www.statcan.gc.ca/eng/start`, you can scroll down to *"Key Indicators"* to find a table showing the components of the CPI.

For April 2018 Statistics Canada reported a CPI of 132.8 compared to a CPI of 100.0 in the base year 2002. That meant the cost of the basket of goods in April 2018 was 32.8 per cent higher than it was in 2002. Prices and the cost of living increased over the 16-year period. At the end of April 2019 the CPI was 135.6. Prices had increased again. **Inflation** is defined as a persistent rise in the general price level as indicated by these increases taking the change, as a percentage, in the price level the previous year.

> **Inflation**: a persistent rise in the general price level.

The inflation rate is calculated using the same method used for calculating the growth rate in real GDP. For example, using CPI values for 2017 and 2018:

$$\text{Inflation rate for 2018} = \frac{\text{CPI}_{2018} - \text{CPI}_{2017}}{\text{CPI}_{2017}} \times 100 \tag{16.2}$$

$$\text{Inflation rate for 2018} = \frac{133.5 - 131.1}{131.1} \times 100 = 1.82\%$$

Statistics Canada also collects and publishes information on the Canadian labour market. It uses a monthly Labour Force Survey of approximately 50,000 Canadian individuals 15 years of age or over living in the provinces of Canada, excluding full-time members of the armed forces, those persons living on Indian reserves, and those in institutions such as penal institutions, hospitals, and nursing homes. The survey provides the data used to estimate the size of the labour force, employment, and unemployment.

Employment is defined as the number of adults (15 years of age and older) employed full-time and part-time and self-employed. **Unemployment** covers those not working but available for and *seeking work*. The civilian **labour force** is those adults who are employed plus those not employed but actively looking for jobs. Based on these concepts, and data on the surveyed population, Statistics Canada reports three key labour market indicators, namely: The participation rate, the unemployment rate, and the employment rate. Employment and unemployment receive most of the media attention and have become familiar indicators of economic conditions. There are, however, two other underlying labour market measures that deserve attention when interpreting the employment and unemployment rates.

> **Labour force**: adults employed plus those not employed but actively looking for work.
>
> **Employment**: number of adults employed full-time and part-time and self-employed.
>
> **Unemployment**: number of adults not working but actively looking for work.

The **participation rate** is the proportion of the surveyed population that is either working or unemployed. It measures the size of the labour force relative to the surveyed population. The participation rate changes as people become more optimistic about finding employment, or discouraged by periods without employment. Discouraged workers want to work but are no longer looking for work because they believe suitable work is not available. As a result they are excluded from the measurement of the labour force and reduce the participation rate. *Changes in the participation rate change the size of the labour force and the unemployment rate even if employment and the population are constant.*

> **Participation rate**: percent of the population that is either working or unemployed.

$$\text{Participation Rate} = \frac{\text{Labour force}}{\text{Population 15+ yrs}} \times 100 \tag{16.3}$$

The **unemployment rate** is the number of unemployed persons expressed as a percentage of the labour force. The size of the labour force depends on the participation rate, which reflects the choices people make about looking for work. The unemployment rate will rise if people become more optimistic about job prospects and begin to look for work, increasing the participation rate and the labour force. On the other hand, the unemployment rate will decline if some people become discouraged and give up looking for work, reducing the participation rate and the labour force.

> **Unemployment rate**: the number of unemployed persons as a percentage of the labour force.

The unemployment rate is calculated as follows:

$$\text{Unemployment Rate} = \frac{\text{Labour force} - \text{employment}}{\text{Labour force}} \times 100 \tag{16.4}$$

Unemployment as measured by the broad unemployment rate has three important components. **Cyclical unemployment** is unemployment that would be eliminated by a higher level of economic activity without putting increased pressure on wage rates and inflation. **Frictional unemployment** comes from the dynamics of the labour market as changing labour force participation and employment opportunities mean that it takes time to match job openings with job candidates. **Structural unemployment** reflects differences in labour force characteristics and employment opportunities as the structure of the economy changes. In combination, frictional and structural unemployment make up the "full employment" level of unemployment. The corresponding unemployment rate is defined as the **natural unemployment rate**. In recent years in Canada, estimates of frictional and structural unemployment suggest a natural unemployment rate of about 6.0 percent. An unemployment rate persistently below 6.0 percent might create inflationary pressure in the labour market and the economy.

Cyclical unemployment: would be eliminated by higher levels of economic activity.

Frictional unemployment: a result of the time involved in adjusting to changing labour force and employment opportunities.

Structural unemployment: caused by changes in economic structure relative to labour characteristics.

Natural unemployment rate: the unemployment rate at "full employment".

Employment rate: percent of the population 15 years of age and over that is employed.

Employment rates provide a different perspective on labour market conditions because they are not affected by changes in the participation rate, which can change unemployment rates. If some people become discouraged and stop looking for work the participation rate, the labour force and the unemployment rate decline, but the employment rate is unchanged. The employment rate is calculated as:

$$\text{Employment Rate} = \frac{\text{Employment}}{\text{Population 15+ yrs}} \times 100 \tag{16.5}$$

Table 16.1 gives data on the Canadian labour force and labour market conditions in 2013 and 2019 in terms of the Participation, Employment and Unemployment rate concepts.

Table 16.1: Change in the Canadian Labour Market, 2013Q1 and 2019Q2 (thousands of persons and percent)

	2013Q1	**2019Q2**	**%Δ**
1. Non-institutional population 15+ yrs	28,504	30,626	7.4
2. Labour force	18,978	20,188	6.4
3. Employment	17,623	19,029	8.0
4. Unemployment	1,355	1,159	-14.5
5. Participation rate $[(2)/(1) \times 100]$	66.7%	65.9%	-
6. Employment rate $[(3)/(1) \times 100]$	61.8%	62.1%	-
7. Unemployment rate $[(4)/(2) \times 100]$	7.2%	5.7%	-

Source: Statistics Canada, Table 14-10-0287-01 and authors' calculations.

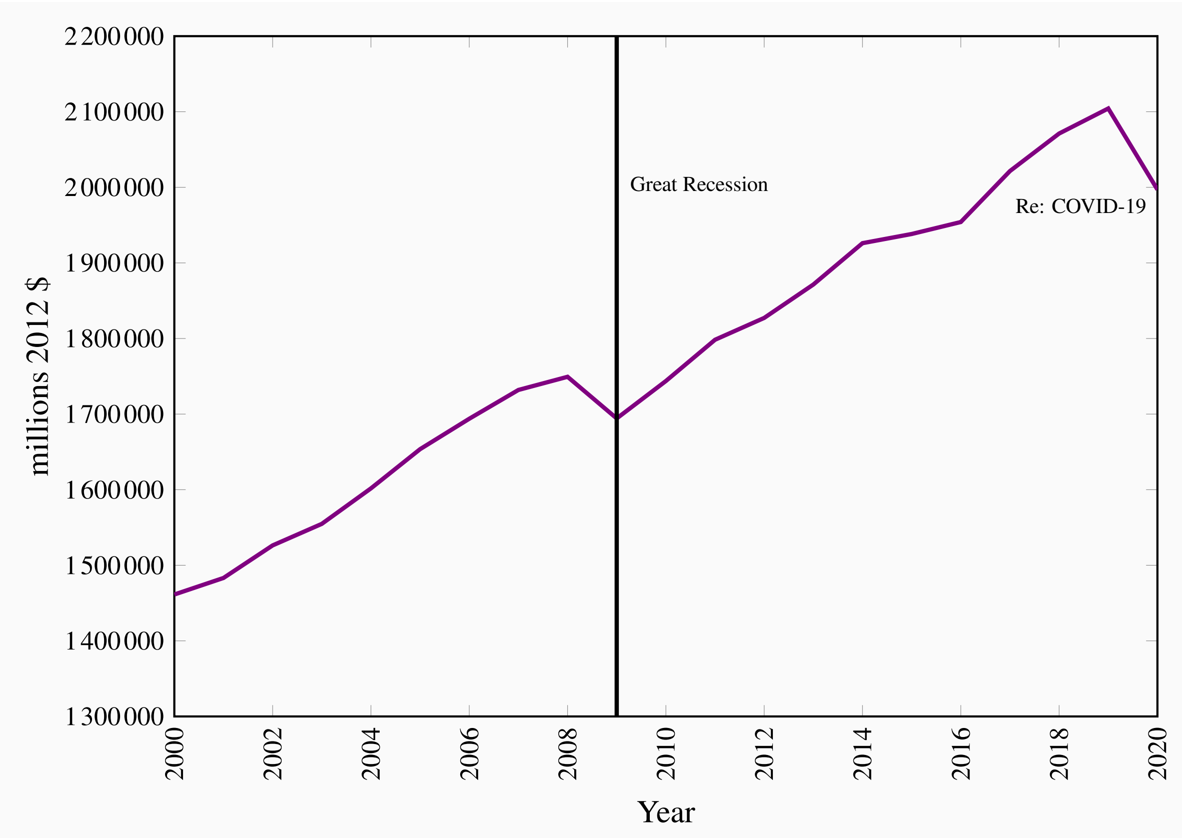
Almost every day the media discuss some aspects of economic growth, inflation, and employment. Often these discussions ignore the requirement that employment must grow faster than the growth in the labour force if unemployment is to decline. Good news about 'job creation' needs to be tempered by news on labour force growth. These issues often play large roles in elections and discussions of economic policy. In the chapters that follow, we will study causes of changes in output, income, prices and inflation, and employment and unemployment. As a background to that work, consider recent Canadian economic performance.

16.2 Recent Canadian economic performance

The positive relationship between economic performance and standards of living motivates the study of macroeconomics and macroeconomic policy. The ideal would be an economy in which, starting with full employment of labour and capital equipment, the rate of growth of real GDP matched the rate of growth of the labour force and growth in labour productivity, at a low and stable rate of inflation. This is sometimes referred to as a 'Goldilocks Economy' – neither too hot nor too cold, just right.

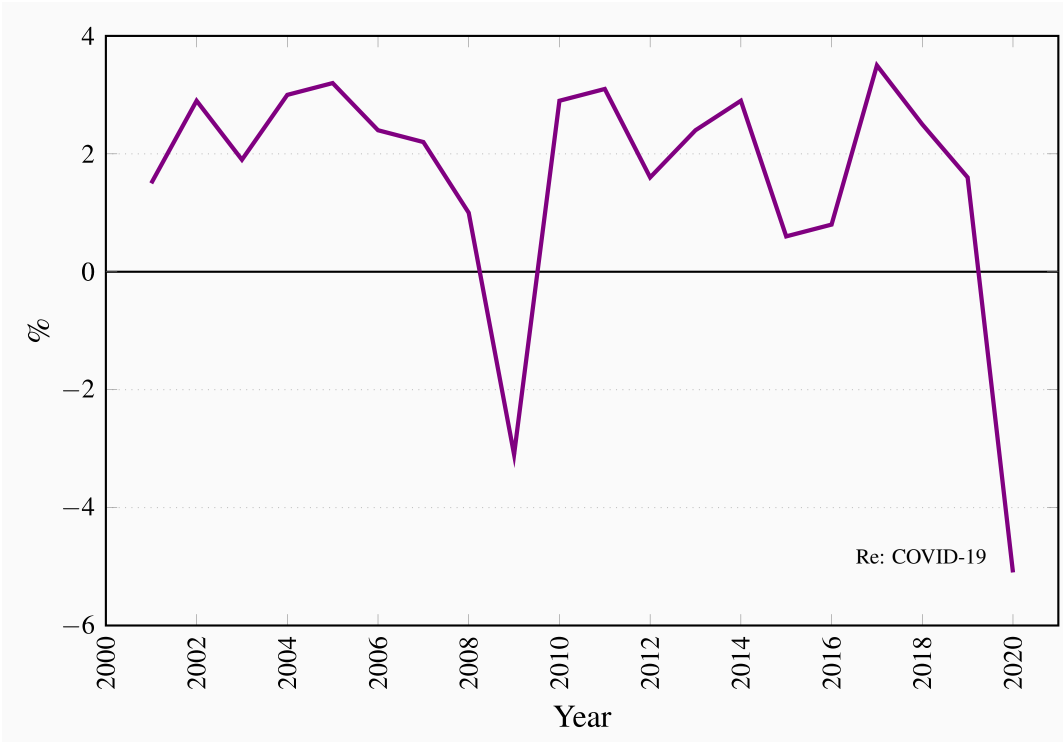
Figures 16.1 to 16.4 provide a more detailed look at the actual rate of growth of real GDP, the rate of inflation, and unemployment rate in Canada over the 2000–2019 time period. They show the trends and annual variations in these measures of economic performance. The observations for 2020 are Parliamentary Budget Office estimates of the effects of policies to control COVID-19. Understanding the causes of these short-term fluctuations in economic performance, their effects on standards of living and the economic policy questions they raise, are major reasons for studying macroeconomics.

Figure 16.1: Real GDP in Canada 2000–2020



Source: Statistics Canada Table 31-10-0369-01 and Parliamentary Budget Office, Report RP 2020-004-s Table 2.1 and Appendix A

Figure 16.2: Annual Real GDP Growth in Canada, 2001–2020



Source: Statistics Canada Table 31-10-0369-01 and Parliamentary Budget Office, Report RP 2021-004-s Table 2.1 and Appendix A

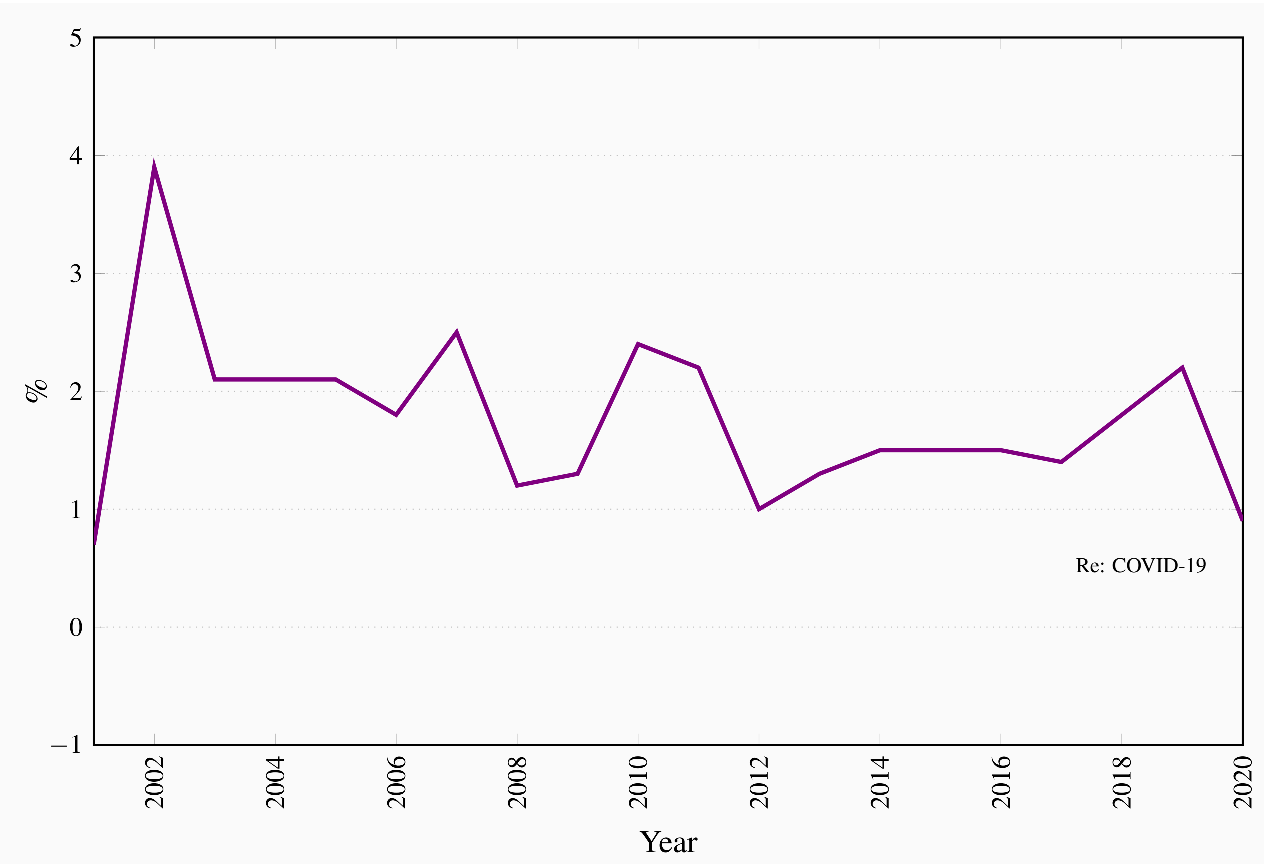
Figure 16.1 shows the substantial growth in real GDP over the 2000–2019 period. It also shows that growth was not steady. Real GDP did increase from 2000–2008 with annual growth rates ranging from 1.5-3.0. Then the real GDP declined sharply by 3.0 percent in 2009. This and other times of negative growth in real GDP are called **recessions**. Indeed the fall in real GDP in 2009, the largest such decline since the 1930s, is now called the 'Great Recession'. The estimated drop in real GDP in 2020, also shown appears in Figure 16.2 as a large negative growth rate. This is the result of businesses closing to enforce physical distancing to control the spread of COVID-19.

Recession: decline in economic activity, often defined as two consecutive quarters of negative growth in real GDP.

Figure 16.2 shows more clearly the considerable fluctuations in real GDP annual growth rates, the negative growth rate of the Great Recession, and the even larger estimate of the negative growth caused by COVID-19. Even when the trend in growth is positive, fluctuations in growth rates can have negative effects on standards of living. They are reflected in changes in employment, changes in incomes and changes in markets that can make life difficult for those affected. We study macroeconomics to find explanations for the causes and effects of these fluctuations in economic activity that will guide stabilization policies.

Figure 16.3 shows annual inflation rates in Canada since 2000. These show the relative stability of Canadian inflation in the years leading up to the Great Recession. That pattern changed after 2008 with the effects of lower output growth and higher unemployment on prices and wage rates. The policy to control COVID-19 will reduce estimated inflation in 2020.

Figure 16.3: Annual Inflation Rates in Canada, 2000–2020 (estimate)
(Annual %Δ in seasonally adj'd December CPI)

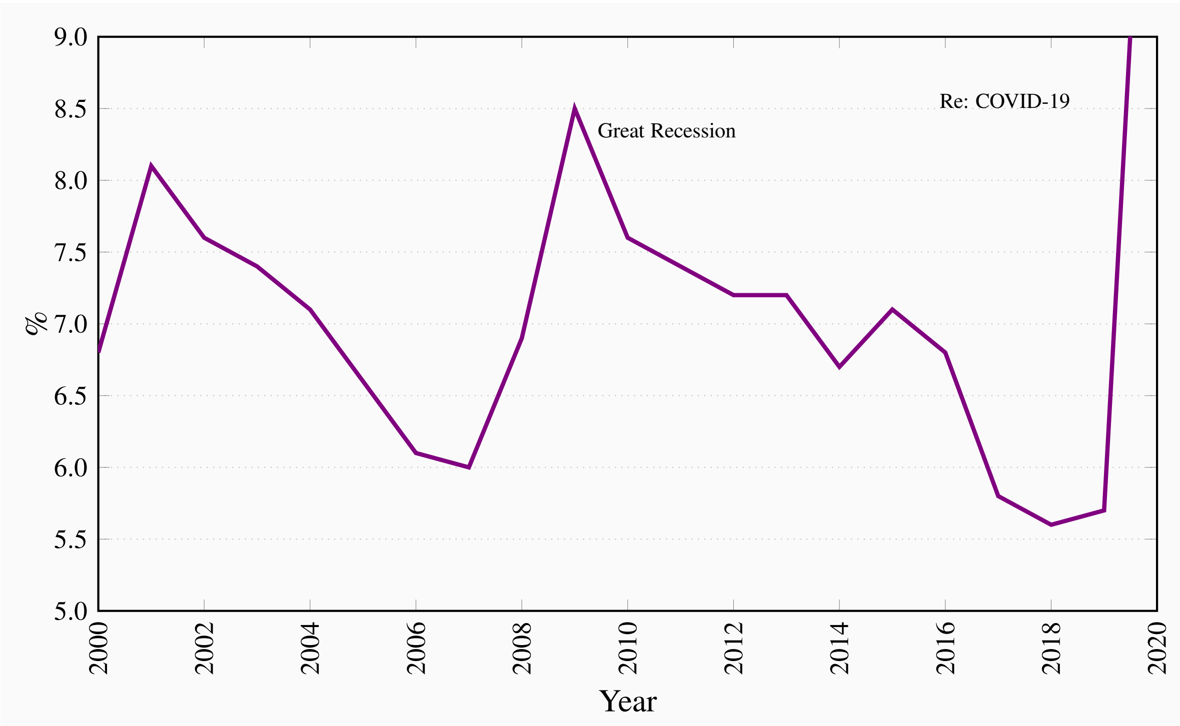


Source: Statistics Canada Table 18-10-0006-01 Consumer Price Index, and Parliamentary Budget Office, Report RP 2021-004-s, Table 2.1 and Appendix A

Our recent experience with low and stable inflation rates in the 2000–2018 period is quite different from past experience. In the late 1980s and early 1990s annual inflation rates were at times higher than 10 percent. We will examine the roles that monetary policies and recessions played in these changes in inflation rates.

Fluctuations in growth rates and inflation rates are also accompanied by fluctuations in unemployment rates. Annual unemployment rates plotted in Figure 16.4 have fluctuated between about 6 percent and 8.5 percent. Although employment has grown over time, when job creation falls short of the growth in the labour force, unemployment rates rise. At other times, strong real GDP growth and job creation have lowered the unemployment rate. The falling unemployment rates from 2002 to 2007 and again from 2009 to 2018 coincided in time with the continuous growth in real GDP we saw in Figure 16.1.

Figure 16.4: Annual Unemployment Rates in Canada 2000–2020



Source: Statistics Canada Table 14-10-0287, and Parliamentary Budget Office Report RP 2021-0004-s, Table 2.1 and Appendix A

The very large increase in the estimate of unemployment in 2020, as a result of COVID-19 policies, overshadows that of the Great Recession. The recovery of GDP growth after the Great Recession was strong enough to offset growth in the labour force and push unemployment rates lower even than pre-recession levels. Economic recovery after COVID-19 may eventually lead to the same result.

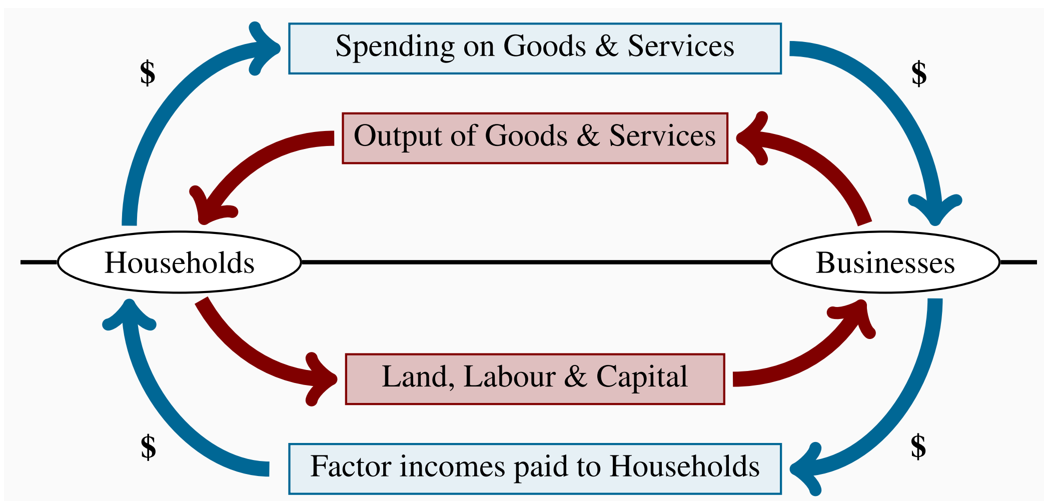
16.3 National accounts and economic structure

National accounts provide the framework that is essential for consistent definitions and measurement of spending, output, and incomes. However, it is important to recognize that this is an accounting system that *describes* the economy, not an economic model that *explains* the economy's behaviour. That comes later.

The national economy involves all households, businesses, other organizations and governments that make decisions about employment, output and expenditures. The results of individual decisions made by these economic units are measured by the economy's total spending, output, and income. The **circular flow diagram** in Figure 16.5 shows the relationship between spending, output, and income.

> **Circular flow diagrams**: show the flows of money payments, real resources, and goods and services between households and businesses.

Figure 16.5: Circular flows in the economy



We start with the simplest of economies. There are only households and businesses; no government and no trade with other countries. Households own the factors of production: Labour, land, capital, and entrepreneurship. Businesses use these factors of production to produce outputs of goods and services. Businesses pay households for the factor services they use and recover these costs by selling their output to the households.

Figure 16.5 shows the circular flow of inputs to production, outputs of goods and services, costs of the inputs to production, and receipts from sales. The upper half of the diagram, above the horizontal line, shows the outputs of goods and services supplied by business to households and household expenditures on those goods and services. The lower half of the diagram shows the factor services of labour, land capital, and entrepreneurship supplied by households to business in exchange for the factor incomes: wages, rent, interest, and profit.

The figure also suggests an alternative way to look at activity in the aggregate economy. The inner loop in the diagram shows the *flows of real factor services* between households and businesses. Households provide factor services to business and get goods and services in return. In modern economies this exchange of factor services for goods and services is facilitated by the use of money as a means of payment. The outer loop in the diagram illustrates the *flows of money payments* made by business to buy factor services, and by households to buy goods and services produced by business. Business pays wages, rent, interest, and profits to households and finances those costs with their receipts from sales to households. To keep the example simple, we assume that households spend all the income they receive from the business sector on goods and services produced by the business sector.

Figure 16.5 illustrates four ways to measure economic activity, namely:

1. the output of goods and services at market prices;
2. the total expenditure on goods and services at market prices;
3. the inputs to the production of goods and services costed at market prices; and
4. the incomes received by households for providing factor inputs to production.

The four rectangles in the diagram show these four alternative but equal measurements.

The *accounting framework* gives the same measure of total economic activity whether we use the market value of output, total spending on that output, inputs to production, or the factor incomes received by households in return for those inputs.

This circular flow model is kept very simple to illustrate the basic accounting principle:

$$\begin{aligned}\text{Market value of output} &= \text{total expenditure}\\ &= \text{market value of factor services}\\ &= \text{household income.}\end{aligned}$$

While the principle illustrated by the circular flow is sound, the economy in Figure 16.5 is too simple. It does not allow households to save or businesses to invest. It leaves out government expenditures and taxes, and transactions between households and businesses with the rest of the world. Including those aspects of economic activity would make our model more complex, and we would need a comprehensive system of national accounts to describe and measure it. But the basic accounting principle would be the same: the four ways to measure total activity in the economy give, *by definition*, the same answer.

16.4 Measuring GDP

Nominal GDP is measured using market prices and a specific time period. It is not possible to add up the final physical outputs of many different businesses and arrive at a meaningful result. Instead, because we have a 'money economy', we let current market prices determine the money values of these different outputs. Then the total market value can be found by adding up the money values. **Nominal GDP** is the market value at current prices of all final goods and services.

> **Nominal GDP**: the market value at current prices of all final goods and services.

Furthermore, the outputs of goods and services occur over time, not all at once. They flow over time and must be measured relative to time. GDP measures over three-month and one-year time periods are reported as quarterly GDP (usually reported at "annual rates") and annual GDP respectively. Annual nominal GDP for any year is the value of the **final goods and services** produced in that year at the prices of that year.

> **Final goods and services**: goods and services are purchased by the ultimate users.

In Canada, Statistics Canada uses the Canadian System of National Accounts (CSNA) to measure GDP. This framework is based on the circular flow concept we have discussed, but is applied to the complexity of the actual economy.

Although earlier in this chapter we defined and discussed *real GDP*, measured at prices of a base year national accounting measures nominal GDP at current prices. The CSNA produces three measurements of nominal GDP:

1. *Output-based GDP* is the sum of value added (output less the cost of goods and services purchased from other business) by all industries in Canada;
2. *Income-based GDP* records the earnings generated by the production of goods and services; and
3. *Expenditure-based GDP* is equal to expenditure on *final goods and services* produced.

These three alternative measures of GDP provide importantly different perspectives on the level of national economic activity. The output and income measures describe the *supply side* of the economy in terms of goods and services produced, and cost of production. The expenditure measure of GDP describes the *demand side* of the economy.

Output-based GDP

To measure output in the economy, and the contribution of particular businesses or industries to that output, we use the **value-added** approach to GDP. Value added measures the *net* output of each industry. To find the value added (net output) of a particular business or industry, the costs of the goods and services purchased from other businesses and industries are deducted from the value of the final product. National, or all-industry GDP, is then the sum of GDP by industry.

> **Value added**: the difference between the market value of the output of the business and the cost of inputs purchased from other businesses.

This method recognizes that businesses buy inputs to production from other businesses as well as from households. Automakers like General Motors and Honda buy parts and components like tires and windshields from other businesses, and include the costs of those inputs in the prices of the finished cars they sell. They also buy services like accounting, advertising, and transportation from service producers. Similarly, pizza makers buy cheese and pepperoni from cheese factories and meat processors. If we were to add up the outputs of auto parts manufacturers, cheese makers, meat processors, pizza makers, General Motors, and Honda in our measurement of nominal GDP, we would overstate GDP by *double counting*. The cheese would be counted once at the cheese factory and again in the pizza. The same applies to the tires and windshields of the new cars. To avoid double counting, we use value added, the increase in the value of goods and services as measured by the difference between market value of output and the cost of **intermediate inputs** bought from other businesses. Or we could count only the outputs sold to *final* users. Notice that total GDP by our definition measures the output of final goods and services.

> **Intermediate inputs**: services, materials, and components purchased from other businesses and used in the production of final goods.

Consider a simple example. A coffee shop sells 100 cups of coffee an hour at a price, before tax, of \$1.50. To make 100 cups of coffee the shop uses 2 kilos of ground coffee costing \$10.00 per kilo, 25 litres of pure spring water costing \$0.40 a litre, and electricity and dairy products costing, in total \$20. The coffee shop's sales per hour are \$150 using inputs costing \$50. Its value added is $\$150 - \$50 = \$100$. As we will see shortly, this value added, or \$100, covers the labour costs, rent, interest expenses, and management costs of the business, for producing 100 cups of coffee an hour.

**Table 16.2: Outputs of Selected Industries in GDP
Canada, July 2018 (percent shares)**

All industries	100.0
Goods producing industries	29.5
Service producing industries	70.5
Agriculture, forestry, fishing, etc.	1.5
Mining, oil and gas extraction	8.3
Construction	7.1
Manufacturing	10.5
Wholesale and retail trade	11.4
Transportation	4.6
Finance, insurance and real estate	19.9
Professional, scientific and management	8.0
Health and social assistance	6.7
Educational services	5.2
Public administration	6.2
Entertainment, recreation, accommodation, etc.	3.0
All other	7.6

Source: Gross domestic product (GDP) at base 2012 prices. Statistics Canada, Table 36-10-0434-01 and author's calculations

Table 16.2 shows the industrial structure of output in Canada in 2018, based on the percentage shares of selected industries in Canadian GDP. Industry outputs are measured by value added. The data illustrate the importance of service-producing industries to economic activity in Canada. This industrial structure is typical of today's high-income economies and raises many interesting questions about the relationship between economic structure, performance, and growth. However, when our main interest is in the total level of economic activity rather than its industrial structure, the expenditure-based and income-based measures of GDP are used.

Expenditure-based GDP and income-based GDP

In the national accounts, expenditure based GDP and income based GDP are equal by definition. Table 16.3 provides an example using the actual accounts for Canada in 2018Q4. Expenditure categories and their shares in total expenditure are recorded in the left side of Table 16.3. Income categories, indirect taxes and their shares in total income are recorded on the right-hand side.

Table 16.3: Canadian National Accounts 2018Q4
($ billions at current prices and seasonally adjusted at annual rates and % GDP)

Expenditure Measures			Incomes Measures		
At market price	$	%	**By income source**	$	%
C by households	1,196.7	53.8	(W) Employee compensation	1,135.0	51.1
I by business	597.6	26.9	(GCS) Gross operating surplus	568.6	25.6
G by government	417.8	18.8	(GMI) Gross mixed income	267.9	12.1
X exports	664.4	29.9	(T_{IN})Net indirect taxes	253.2	11.4
IM imports	−654.8	−29.5			
Statistical discrepancy	1.4	0.1	Statistical discrepancy	-1.5	-0.1
GDP at market price	**2,223.1**	100.0	**GDP at market price**	**2,223.2**	100.0

Source: Statistics Canada, Tables 36-10-0103-01 and 36-10-0104-01 and author's calculations

Expenditure-based nominal GDP adds up the market value of all the final goods and services bought in a given time period, say one year. The national accounts classify this final expenditure into five main categories: Consumption, investment, government expenditure, exports, and imports.

For expenditure, the national accounts classification system is essential for our study of macroeconomic activity for two reasons. First, the classification scheme covers final expenditure in the economy completely; nothing is omitted. Second, the categories represent expenditure decisions made for different reasons in different parts of the economy and the percentage share or importance of each in final expenditure. Understanding expenditure decisions is critical to the work that lies ahead. Defining the *expenditure* categories is the first step.

Applying a name to each expenditure category in the table and using the notation attached gives:

$$\text{GDP} = \text{consumption} + \text{investment} + \text{government expenditure} + \text{exports} - \text{imports}$$

or

$$\text{GDP} = C + I + G + X - IM \tag{16.6}$$

For macroeconomic theory and models, this expenditure GDP is the foundation of the theory of aggregate demand introduced in Chapter 17 and developed in detail in later chapters.

Income-based GDP adds up the factor costs of production of all goods and services plus the net indirect taxes included in market price. The table also shows the percentage share of each category in GDP and thus the relative importance of each in income or cost respectively. The income categories are: Employment compensation (W), gross operating surplus (corporate profit) (GCS), gross mixed income (unincorporated business income plus investment income) (GMI) and net indirect taxes (T_{IN}). (The allowance for depreciation of capital is included in GCS and GMI.) Then GDP at market price is:

$$\text{GDP} = W + GCS + GMI + T_{IN} \tag{16.7}$$

This income based GDP measures total cost of production. The first three components W, GCS and GMI are factor costs of production including the depreciation of capital equipment used in production. Net indirect tax T_{IN} is the revenue generated by taxes applied to goods and services and included in final price. An *aggregate supply function* for the economy involves these costs of production in relation to total output. Aggregate expenditure at market prices is the revenue that producers receive to cover these costs.

To construct a macroeconomic theory and model of the economy we must explain the linkages, feedbacks and interactions among the elements of the economy defined by national accounting conventions. These linkages, feedbacks and interactions are the important relationships that work together to explain how this economic system determines GDP, business cycle fluctuations in GDP, inflation, and employment.

16.5 Nominal GDP, real GDP & the GDP deflator

We have used *real GDP* to measure growth and the growth rate in the beginning of this chapter, and then *nominal GDP* as recorded in the National Accounts. Now we need to look carefully at both concepts and the relationship between them, which is the GDP deflator, a measure of the general price level. *Nominal GDP* measures output and incomes based on *current market prices* for goods and services and factors of production. As a result, changes in nominal GDP from one period to the next might be the result of changes in prices of final outputs and factor inputs, or the result of changes in the quantities of final outputs and factor inputs, or some combination of the two. Since it is physical quantities of goods and services that yield satisfaction or utility, it can be misleading to judge the economy's performance by looking at nominal GDP. For that purpose we need real GDP, as we discussed earlier in this chapter. *Real GDP*, or *GDP in constant prices*, measures the value of goods and services produced in any given year using the prices of a base year. In this way, real GDP adjusts changes in GDP for changes in prices by measuring GDP in different years in constant prices.

Table 16.4: Nominal and real GDP

		2008	2018	% change
Quantity	blue jeans	4,000	5,000	25
	solar panels	2,000	4,000	100
Price in current $	blue jeans	25	50	100
	solar panels	100	60	−40
Current value	blue jeans	100,000	250,000	150
	solar panels	200,000	240,000	20
Nominal GDP		**300,000**	**490,000**	**63**
Value in 2007 $	blue jeans	100,000	125,000	25
	solar panels	200,000	400,000	100
Real GDP		**300,000**	**525,000**	**75**
GDP deflator		**100**	**93.3**	**−6.7**

To illustrate this important point, Table 16.4 shows a simple economy that produces both consumer goods, blue jeans, and capital goods, solar panels. In this economy nominal GDP rises from $300,000 to $525,000 between 2008 and 2018, a 63 percent increase measured in current prices as a result of changes in both quantities and prices. If we take 2008 as the base year, we can measure real GDP in 2018 by valuing output quantities in 2018 using 2008 prices. This gives real GDP in 2018 of $525,000 in prices of the base year. In the example in the table, quantities of both products rise over the period ***but the price of blue jeans rises while the price of solar panels falls***. As a result the rise of about 75 percent in real GDP gives a true picture of the extra quantity of goods available in the economy in 2018 compared with 2008. It eliminates the change in nominal GDP that was the result of the fall in the average price level by 6.7 percent between 2008 and 2018 as a result of the fall in the price of solar panels.

The GDP deflator

The Canadian economy is obviously more complex than this economy. We have seen that GDP includes expenditures by households, governments, businesses, and residents of other countries who supply us with imports and buy our exports. To convert nominal GDP to real GDP we need to use an index that includes what is happening to the prices of all these different goods and services. This index is called the **GDP deflator**.

> **GDP deflator**: index of current final output prices relative to base year prices.

If we have data for both nominal and real GDP, we can calculate the GDP deflator as the ratio of nominal GDP to real GDP expressed as an index with a value of 100 in the base year.

$$\text{GDP deflator} = \frac{\text{Nominal GDP}}{\text{Real GDP}} \times 100 \tag{16.8}$$

The GDP deflator differs from the consumer price index (CPI) illustrated in Example Box 16.1 and used to measure inflation in consumer prices and the cost of living. First, the CPI is based on a "representative basket" of goods and services that consumers buy, while the GDP deflator is comprehensive and covers all the goods and services included in national accounts. Second, the CPI changes over time with changes in the prices of the basket of consumer goods and services. The GDP deflator, by contrast, is built on the base year prices. It changes over time as the current prices change relative to base year prices. In other words the GDP deflator is used to "deflate" the dollar value of current 2018 output to what value it would be in 2008 prices, while the CPI measures the *increase in the cost* of the "basket" of consumer goods and services.

But why does the GDP deflator change over time? From our earlier discussion of the national income accounting framework, we can see that *costs of production* and *net indirect taxes* are included in the general level of market prices measured by the GDP deflator. Nominal GDP measured by the income approach is reported in Table 16.3. It is the sum of incomes paid to factor inputs to production, plus depreciation allowances and net indirect taxes. These components of nominal GDP are the costs of production, gross profits, and taxes that are built into the market prices of the goods and services.

The general price level in the economy is the dollar amount paid for a 'unit of output' and, subtracting indirect taxes, the revenue received by producers for the sale of 'unit of output'. Revenue per unit of output must cover costs per unit of output, including expected profit, for producers to be willing to continue operations. Changes in costs must be covered eventually by changes in prices. Or if market conditions raise prices—think about crude oil production or lumber production—producers will increase output, as long as higher prices cover higher costs.

Summarizing from the national accounts gives three components of cost per unit of output:

1. employee compensation per unit of output, W/Y;
2. gross business income per unit of output, $(GCS+GMI)/Y$; and
3. net indirect tax per unit of output, T_{IN}/Y.

Changes in the sum of these three components of the price level must change *both* price and nominal GDP, whether we measure nominal GDP by the income or the expenditure approach. The GDP deflator is an index of this price level in any particular year relative to a chosen base year. However, the accounting framework does not explain the causes of change in the price level. That requires explanations of changes in unit labour costs, of producer output and pricing decisions and information on the net indirect tax rate. Those explanations are parts of an economic model of the supply side of the economy.

To show the empirical importance of the distinction between real and nominal GDP, Table 16.5

gives Canadian data over the period 2006 to 2018. Nominal GDP rose from $1,510 billion in 2006 to $2,225 billion in 2018. Without knowing what happened to prices of goods and services in general, we cannot judge what happened to the quantity of output over that period. To answer this question we use the GDP deflator to convert nominal GDP to real GDP in the prices of the base year 2007 as follows:

$$\text{Real GDP}_{\text{year } t} = \frac{\text{GDP}_{\text{year } t}}{\text{GDP deflator}} \times 100 \tag{16.9}$$

Table 16.5: Canadian nominal and real GDP 2006Q4–2018Q2 (seasonally adjusted at annual rates)

	2006Q4	2010Q4	2014Q4	2018Q2
Nominal GDP (billions $)	1,510	1,698	2,009	2,225
GDP deflator (2007=100)	98.4	102.4	112.9	120.2
Real GDP (billions 2007 $)	1,534	1,658	1,779	1,851

Source: Statistics Canada Tables 36-10-0104-01, 36-10-0105-01 and 36-10-0106-01

For example, in 2018Q2, nominal GDP was $2,225 billion and the GDP deflator ($2007 = 100$) was 120.2. Real GDP measured in constant 2007 dollars was then:

$$\text{Real GDP}_{2018Q2} = \frac{\$2,225}{120.2} \times 100 = \$1,851 \text{ in 2007 dollars}$$

When converted to constant dollars, the change in real GDP is much smaller than the change in nominal GDP. Over the 2006–2018 period shown in the table, real GDP increased by 20.7 percent compared to a 47.4 percent increase in nominal GDP. On average, prices in 2018 were 22.2 percent higher than in 2006. Clearly, it is important to distinguish between nominal and real GDP.

16.6 Per capita real GDP and standards of living

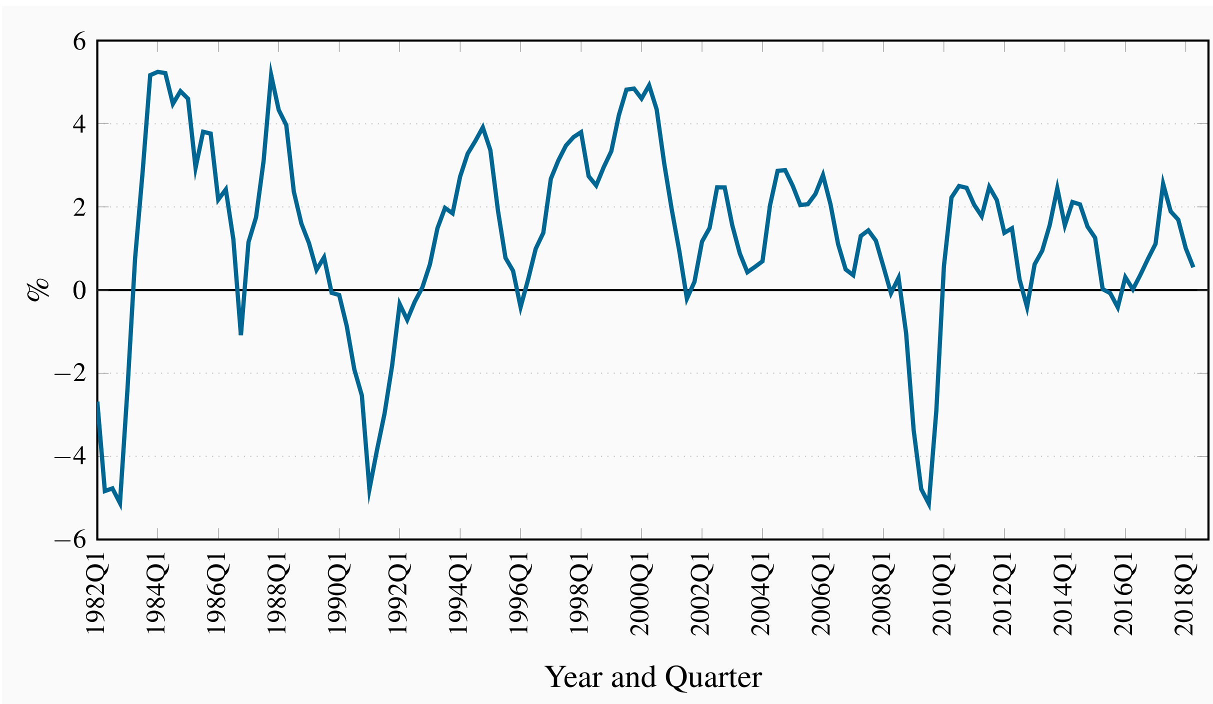
Real GDP is a simple measure of the total real income and output of an economy. The percentage change in real GDP we saw in Figures 16.1, 16.2 and 16.3 shows how fast the economy is growing. But we are also interested in what is happening to *productivity* and the *standard of living* in the economy and how they change over time. For a given real GDP, the larger the population, the lower is productivity and the smaller is the quantity of goods and services per person. To get a simple measure of the standard of living enjoyed by a person in the economy it is better to look at **per capita real GDP**, which adjusts for population. Whether or not growth in total GDP improves standards of living depends also on what is happening to the size of the population. To find per capita real GDP for a country, which is real GDP per person, we simply divide real GDP by population.

Per capita real GDP: real GDP per person.

$$\text{Per capita real GDP} = \frac{\text{Real GDP}}{\text{Population}} \tag{16.10}$$

The study of short-run macroeconomics is strongly motivated by the negative effects of recessions on national standards of living. Figure 16.6 shows the negative effects of recessions of 1982, 1991, and 2009 on per capita GDP.

Figure 16.6: Quarterly Growth in Real GDP Per Capita at annual rates Canada, 1982–2018



Source: Statistics Canada Tables 17-10-0009-01 and 36-10-0104-01 and author's calculations

Macroeconomic models are built to help us understand the causes of fluctuations in real GDP, employment, and the price level. Understanding the workings of the economy is essential for the design and implementation of monetary and fiscal policies that contribute to economic stability and protect standards of living.

In longer time horizons macroeconomics seeks to understand and explain the growth of real GDP that is essential to protect and improve standards of living as population grows. Growth also increases the capacity of the economy to direct its resources to a wider range of activities that may include improvements in the quality of goods and services produced or reductions in the effects of growth on social and environmental conditions.

Limitations of real GDP

Because we use GDP to measure the output and income of an economy, the coverage should be as comprehensive as possible. We should also recognize that the composition of GDP and the distribution of income are important to a country's standard of living.

In practice, we encounter several problems when including all production in GDP. First, some production causes noise, pollution, and congestion, which do not contribute to economic welfare. Current national and international concern about carbon dioxide and climate change is a clear and obvious example of the issues involved. We should adjust GDP for these costs to evaluate standards of living more accurately. This is sensible but difficult to do. Recent policy changes by governments to impose carbon taxes on fuels and fuel efficiency targets for automobiles aim to reduce some greenhouse gases. But most such nuisance goods are not traded through markets, so it is hard to quantify their output or decide how to value their costs to society.

Similarly, many valuable goods and services are excluded from GDP because they are not marketed and therefore are hard to measure. These include the home cleaning, maintenance, and improvements households carry out for themselves, and any unreported jobs and incomes in the economy. Deducting nuisance outputs and adding the value of unreported and non-marketed incomes would make GDP a more accurate measure of the economy's production of goods and services.

Furthermore, high GDP and even high per capita GDP are not necessarily good measures of economic well-being. The composition of that output also affects standards of living. Health care services are likely to have different effects than military expenditures. The United Nations prepares an annual *Human Development Index* (HDI) to provide a more comprehensive measure of a country's achievements. The HDI provides a summary measure based on life expectancy, adult literacy, and real GDP per capita.

Table 16.6 shows HDIs for the top ten countries in 2017, according to the *Human Development Report, 2018*. The second last and last columns in the table are of particular interest. The second last column shows the HDI adjusted for national per capita GNP. The underlying argument is that ranking national economic well-being simply by using per capita GNP would miss the importance of life expectancy and education as indicators of standards of living. For example, Singapore and the United States would rank highest by per capita GNP alone, but that ranking is reduced by lower life expectancy and years of schooling. The last two columns in the table compare HDI rankings in 2017 compared to 2016.

Table 16.6: Top twelve countries based on the United Nations human development index

HDI Rank	Country	HDI	Life Expectancy (years)	Expected yrs schooling	Mean yrs schooling	GNP/pop[1]	GNP/pop – HDI rank[2]	HDI rank
		2017	**2017**	**2017**	**2017**	**2017**	**2017**	**2016**
1	Norway	0.953	82.3	17.9	12.6	68,012	5	1
2	Switzerland	0.944	83.5	16.2	13.4	57,625	8	2
3	Australia	0.939	83.1	22.9	12.9	43,560	18	3
4	Ireland	0.938	81.6	19.6	12.5	53,754	8	4
5	Germany	0.936	81.2	17.0	14.1	46,136	13	4
6	Iceland	0.935	82.9	19.3	12.4	45,810	13	6
7	Sweden	0.933	82.6	17.6	12.4	47,766	9	7
8	Singapore	0.932	83.2	16.2	11.5	82,503	-6	8
9	Netherlands	0.931	82.0	18.0	12.2	47,900	5	10
10	Denmark	0.929	80.9	19.1	12.6	47,918	3	10
11	Canada	0.926	82.5	16.4	13.3	43,433	10	12
12	United States	0.924	79.5	16.5	13.4	54,941	-2	12

Source: United Nations Human Development Reports: 2018 update HDR Report.
http://hdr.undp.org/en/composite/HDI

Note[1]: GNP in PPP international dollars.

Note[2]: Difference between rank by GNP per capita and by HDI value.
A negative value means the country is better ranked by GDP than HDI.

Conclusion

In this chapter we have looked at indicators of macroeconomic activity and performance, and the measurement of macroeconomic activity using the national accounts. We have not examined the conditions that determine the level of economic activity and fluctuations in that level. An economic model is required for that work. In the next chapter we introduce the framework of a basic macroeconomic model.

Key Concepts

Macroeconomics studies the whole national economy as a system. It examines expenditure decisions by households, businesses, and governments, and the total flows of goods and services produced and incomes earned.

Real Gross Domestic Product (GDP), prices and **inflation rates**, and employment and **unemployment rates** are indicators of macroeconomic activity and performance.

Fluctuations in the growth rate of real GDP, in inflation rates, and in unemployment rates are important aspects of recent economic performance in Canada.

Circular flow diagram: illustrates the flow of real resources and money payments in terms of the expenditures of households, production of goods and services by business, and the incomes that result, earned in the economy.

The **National Accounts** provide a framework for the measurement of the output of the economy and the incomes earned in the economy.

Nominal GDP measures the output of final goods and services at *market prices* in the economy, and the money incomes earned by the factors of production.

Real GDP measures the output of final goods and services produced, and incomes earned at *constant prices*.

The **GDP deflator** is a measure of the price level for all final goods and services in the economy.

Real GDP and **per capita real GDP** are crude measures of national and individual welfare. They ignore non-market activities, the composition of output, and the distribution of income among industries and households.

Exercises for Chapter 16

Exercise 16.1 You have the following annual data for an economy:

Year	Real GDP (2009 $)	Consumer price index (2009=100)	Labour force (000)	Employment (000)
2016	1,282	109.1	17.593	16.537
2017	1,307	111.9	17.857	16.696
2018	1,288	138.9	18.125	16.856

(a) What was the rate of growth of real GDP from 2016 to 2017, and 2017 to 2018?

(b) What was the rate of inflation in 2017 and in 2018?

(c) What were the rates of growth of the labour force and employment from 2016 to 2017, and 2017 to 2018?

(d) What happened to the unemployment rate between 2016 and 2017, and between 2017 and 2018?

Exercise 16.2 Suppose the economy represented by the table in Exercise 16.1 above had a population of 27.885 thousand in 2018.

(a) What were the participation and employment rates in the economy in that year?

(b) Suppose a mild recession in that year discouraged some unemployed workers and they stop looking for work. As a result the participation rate fell to 64.5 per cent. How would the unemployment rate and the employment rate be affected? Why?

Exercise 16.3 If brewers buy barley and hops from agricultural producers, natural gas to fire their brew kettles from gas companies and bottles from glass manufacturers as in the following table, what is the value added of the brewing industry? If brewers also wholesale some of their output to pubs, is that output counted in GDP? Explain your answer.

	Costs (millions of current $) of:		
Brewery retail sales	Barley and hops	Natural gas	Bottles
1000	350	125	150

Exercise 16.4 The economy has two main industries. One produces services and the other produces goods. The services industries produce services for households and businesses with a total

market value of \$10,000. The goods industries produce goods for the use of both households and businesses with a total market value of \$5,000. The service industries spend \$1,000 on computers and paper and envelopes supplied by the goods industries. The goods industries spend \$1,000 to buy financial, insurance, advertising and custodial services supplied by the service industries. Explain how you measure nominal GDP in this economy and the value of output you find?

Exercise 16.5 Suppose you are given the following data on incomes and expenditures for the economy of Westland, in current prices for factors of production and outputs.

Consumption expenditures	2,500
Employment compensation	2,800
Government expenditure	800
Net indirect taxes	150
Exports	1,200
Gross corporate surplus and mixed income	1,050
Investment expenditure	600
Imports	1,100

(a) What is the value of nominal GDP measured by expenditures?

(b) What is net domestic income?

(c) What is the value of nominal GDP measured by the income approach?

Exercise 16.6 Suppose GDP is \$2,000, consumption expenditure is \$1,700, government expenditure is \$50, and net exports are \$40.

(a) What is business investment expenditure?

(b) If exports are \$350, what are imports?

(c) In this example, net exports are positive. Could they be negative?

Exercise 16.7 Consider the following information about a hypothetical economy:

Year	**Nominal GDP (billions \$)**	**GDP deflator** (2000 = 100)	**Population (millions \$)**
2017	750	104.0	25.0
2018	825	112.0	30.0

(a) Calculate the growth (percentage change) in nominal GDP from 2017 to 2018.

(b) What was real GDP in 2017 and 2018? How much did real GDP grow?

(c) If changes in the standard of living can be measured by changes in real per capita GDP, did growth in nominal and real GDP raise the standard of living in this economy from 2017 to 2018?

(d) Explain the reasons for the change in standard of living that you have found.

Chapter 17

Output, business cycles, growth & employment

In this chapter we will explore:

Many economic events seem simple and limited in their effects on rates of growth of output, rates of inflation and rates of unemployment. A housing market collapse in the US in 2008 was seen, at first as just a collapse in prices in an overheated market. Sharp drops in crude oil prices in early 2015 and in 2019 were seen as a benefit to households that would reduce their driving costs and allow for higher expenditure in other areas, but caused economic distress for Alberta, an oil producing province. Fiscal austerity aimed at achieving balanced government budgets in Europe seemed an obvious way to reduce high government debt to GDP ratios. But none of these initial judgments worked out. In each case, ignoring the complexity of the macro economy led to large errors in early forecasts and persistent, widespread problems with economic performance.

An aggregate demand and aggregate supply model is the workhorse of macroeconomics. It integrates the effects of economic disturbances, economic decisions, relationships and linkages that determine real GDP and the GDP deflator. Analyzing events like financial shocks, commodity price shocks and government policy shifts using a full macroeconomic model provides a better understanding of their effects on rates of growth of GDP, rates of unemployment and rates of inflation. In this chapter we introduce a basic aggregate demand and supply model and use it to illustrate business cycle fluctuations in real output and prices. The economic theory and policy on which aggregate demand and supply are based are developed in the chapters that follow.

Under 'normal conditions' the basic AD/AS model offers explanations of changes in economy wide economic performance resulting from a disturbance or shock to just one sector of the economy. That might be a change in AD as a result of a change in business investment expenditure or exports when AS conditions are stable. Or it might be the effect of change in AS as a result of a change in the wage rate or a change in technology that changes productivity when AD conditions are stable. Or it might be the effect of change in government expenditure or monetary policy. In any of these cases only one of either AD or AS shifts. The other does not. New finite equilibrium values on real GDP and the price level P are predicted by the new intersection of the curve that has shifted with that one remains unchanged.

Large random shocks external to the economy, and government policy responses, complicate the use of the basic AD/AS model. In Canada and other countries, business shutdowns and social distancing policies used to control the spread of COVID-19 epidemic are examples. The shutdowns reduce output, AS, and employment. The unemployment, and losses of business and household income, reduce AD even with substantial government income support and monetary expansion. AD and AS curves both shift to the left as the economy adjusts to a new equilibrium real GDP and price level. But the size of the effects on national income and the prices are not easily determined.

The last section of this chapter looks at short term estimates of the effects of COVID-19 policies on Canadian real GDP growth rates, unemployment rates and inflation rates. The preceding sections examine economic performance before COVID-19.

17.1 Aggregate demand & aggregate supply

The **short run** in macroeconomics is defined by assuming a specific set of conditions in the economy. These are:

1. There are constant prices for factors of production, especially money wage rates for labour.
2. The supply of labour, the stock of capital, and the state of technology are fixed.

> **Short run**: a time frame in which factor prices, supplies of factors of production, and technology are fixed by assumption.

In the short run, changes in output involve changes in the employment of labour and in the use of plant and equipment, but these changes may not be sustainable over longer time periods. Furthermore, because supplies of factor inputs and technology are fixed, there is no sustained growth in real GDP. We leave that topic for a later chapter.

The national accounts we studied in Chapter 16 describe economic activity in terms of an accounting framework used to measure aggregate expenditures, outputs, and incomes. But the accounting framework simply measures what has happened in the recent past. It does not *explain* the level of economic activity and prices or the reasons for changes in output and prices from time to time.

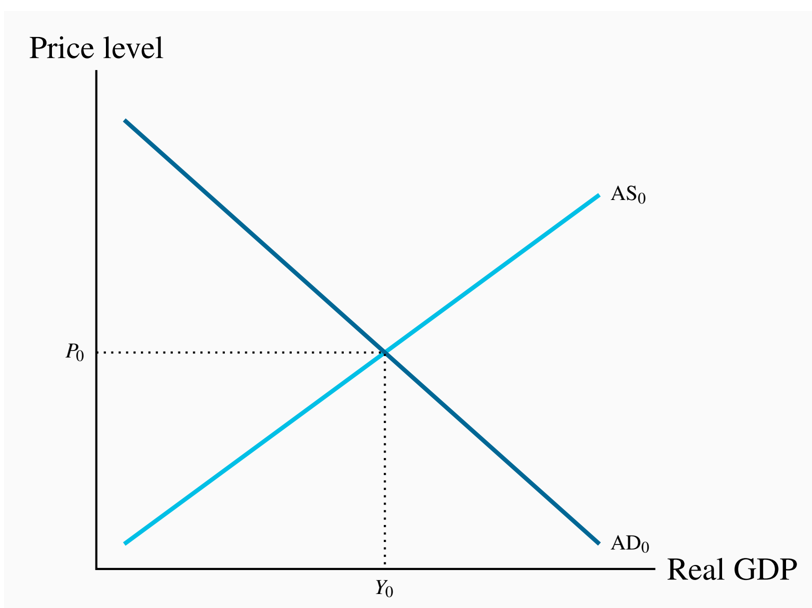
For that we need an analytical framework that looks at cause and effect. An **aggregate demand** and **aggregate supply (AD/AS) model** is such an analytical framework. It helps us understand the conditions that determine output and prices, and changes in output and prices over time.

> **AD/AS model**: a framework used to explain the behaviour of real output and prices in the national economy.

The short-run AD/AS model builds on the national accounts framework. Aggregate demand is the relationship between aggregate expenditure on final goods and services and the general price level. Real GDP by the expenditure approach measures this expenditure at the price level given by the GDP deflator. Aggregate supply is the relationship between the output of goods and services produced by business and the general price level. Real GDP by the income approach measures this output, and the corresponding real incomes. The price level is again the GDP deflator. National accounts tell us that, by definition, these *measured* outputs and incomes are equal. AD and AS functions describe *expenditure plans, outputs, and prices* using the national accounts framework. This distinction between measured and planned expenditure and output is important. Planned expenditure is the current output households and businesses would *want to buy* at different levels of income and price. Output is what businesses actually produce. Planned expenditure and the actual output produced by business may not be the same.

Figure 17.1 gives us a first look at output, real income, and prices for a specific year using an aggregate demand and aggregate supply diagram. The price level as measured by the GDP deflator is measured on the vertical axis. Real output and income are measured on the horizontal axis. The point of intersection of the AD and AS lines shows that real output by the expenditure approach, Y_0, is equal to real income by the income approach at the price level P_0, as required by national accounts. It also shows planned aggregate expenditures equal to the current output of goods and services. However, we need to explain the aggregate demand and aggregate supply relationships indicated by the slopes and positions of the AD and AS lines in the diagram before we use the model to study output and prices.

Figure 17.1: A basic aggregate demand and supply model



The AD and AS lines show planned expenditures on, and the output of, final goods and services at different aggregate price levels *all other conditions held constant*. At the intersection of AD_0 and AS_0 planned expenditures on final goods and services are equal to real GDP at P_0.

> **Equilibrium GDP**: AD=AS, planned expenditure equals current output and provides business revenues that cover current costs including expected profit.

Aggregate demand (AD) is **planned aggregate expenditure** on final goods and services at different price levels when all other conditions are constant. This relationship is examined in detail in the chapters that follow. A downward sloping AD curve means the relationship between planned aggregate expenditure and the general price level is negative. A higher price level reduces the expenditures planned by households, businesses, and residents of other countries. Lower price levels increase those expenditure plans.

> **Aggregate demand**: **planned aggregate expenditure** on final goods and services at different price levels, all other conditions remaining constant.

Aggregate Supply (AS) is the output of final goods and services business produces at different price levels when other conditions are constant. The upward sloping AS curve in Figure 17.1 assumes that the relationship between the quantity of goods and services produced and the price level is positive. Prices and output rise or fall together. We will examine this relationship in more detail below and in later chapters.

As we can see in the diagram, changes in either AD or AS would result in changes in the point of intersection of AD and AS and changes in equilibrium P and Y. But the important question is: What economic conditions and events determine the positions and slopes of AD and AS? The model is a tool for economic analysis that will only be useful when we know how it works and how to operate it.

> **Aggregate supply**: the output of final goods and services businesses would produce at different price levels, all other conditions held constant.

Aggregate demand

Aggregate Demand and the market demand for an individual product are different. In our discussion of the market for an individual product in Chapter 3, demand is based on the assumptions that incomes and prices of other products are constant. Then a rise in the price of the product makes the product more expensive relative to income and relative to other products. As a result, people buy less of the product. Alternatively, if price falls people buy more.

The link between the *general* price level and aggregate demand is different. We cannot assume constant incomes and prices of other products. In the aggregate economy a rise in the price level raises money incomes by an equal amount. A 10 percent rise in the general price level is also a 10 percent rise in money incomes. Changes in the price level do not make goods and services either more or less affordable, *in terms of incomes*. There is no direct price incentive to change aggregate expenditure.

Furthermore, if prices of individual goods and services do not rise or fall in the same proportion as the general price level, the distribution of aggregate expenditure among goods and services may change without a change in aggregate expenditure. If, for example, the general price level is pushed up because oil and commodity prices rise, and expenditure on those products rises in the short run because there are no alternatives, expenditures on other goods and services fall. Aggregate expenditure is unchanged.

As a result, we cannot explain the negative relationship between the general price level and aggregate expenditure as we would explain demand for an individual good or service. Nor can we simply add up all the demands for individual products and services to get aggregate demand. The assumptions of constant incomes and other product prices that underlie market demand do not hold in the aggregate. Different explanations are needed.

Money and financial markets play key roles in the explanation of the price-quantity relationship in aggregate demand and the negative *slope of AD* as follows:

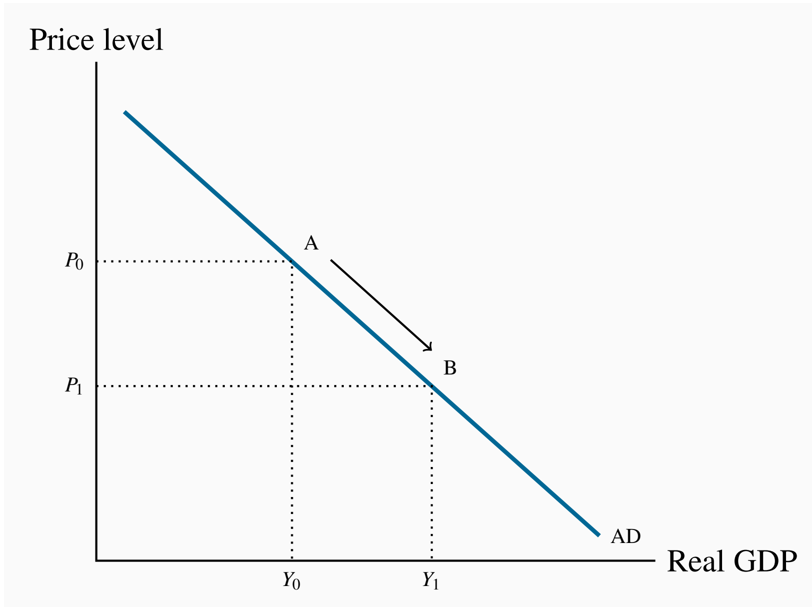
- Changes in the price level (P) change demand and supply conditions in financial markets. Higher prices raise interest rates and lower prices lower interest rates.
- Interest rates determine costs of credit and foreign exchange rates. Higher interest rates reduce expenditures on goods and services, lower interest rates stimulate expenditure.
- The responsiveness of expenditures to changes in interest rates determines the extent of the change in expenditure (Y) as a result of a change in the price level (P).
- Changes in expenditure then feed back to offset some of the change in financial conditions.

Chapters 20, 21 and 22 examine these financial markets and their effects on expenditure.

The result is a negative relationship between the price level and aggregate expenditure and a negatively sloped AD curve.

In Figure 17.2, the negatively sloped AD line shows planned aggregate expenditures at different price levels, on the assumption that the money supply and anything other than price that might affect expenditure plans are held constant. If the price level falls from P_0 to P_1, the movement along AD from A to B shows the negative relationship between planned aggregate expenditure and price. A rise in price would reduce planned expenditure as shown by moving up the AD curve.

Figure 17.2: The aggregate demand curve



The AD curve shows planned expenditures at different aggregate price levels all things other than price held constant. *A change in the price level causes movement along the AD curve* as from A to B if price falls from P_0 to P_1.

The *position of the AD curve* depends on all the conditions other than price that affect aggregate expenditure plans. We study these other conditions in detail in later Chapters 18 and 19.

Aggregate supply

Aggregate Supply (AS) is the output of final goods and services businesses would produce at different price levels. The aggregate supply curve is based on the following key assumptions:

1. Prices of the factors of production—the money wage rate for labour in particular—are constant.
2. The stock of capital equipment—the buildings and equipment used in the production process—and the technology of production are constant.

From national accounts we know that the costs of production include labour costs, business and investment incomes and depreciation. Market prices depend on those *costs per unit of output* and the output and price setting decisions by producers. Aggregate supply is usually described as a positive relationship between quantities of goods and services businesses are willing to produce and prices. Higher outputs of final goods and services and higher prices go together.

This relationship between aggregate output, costs and prices reflects two different market conditions on the supply side. In some markets, particularly those for commodities and standardized products, supply and demand in international markets establish price. Producers of those products are price takers. They decide how much labour and plant capacity to employ to produce based on market price.

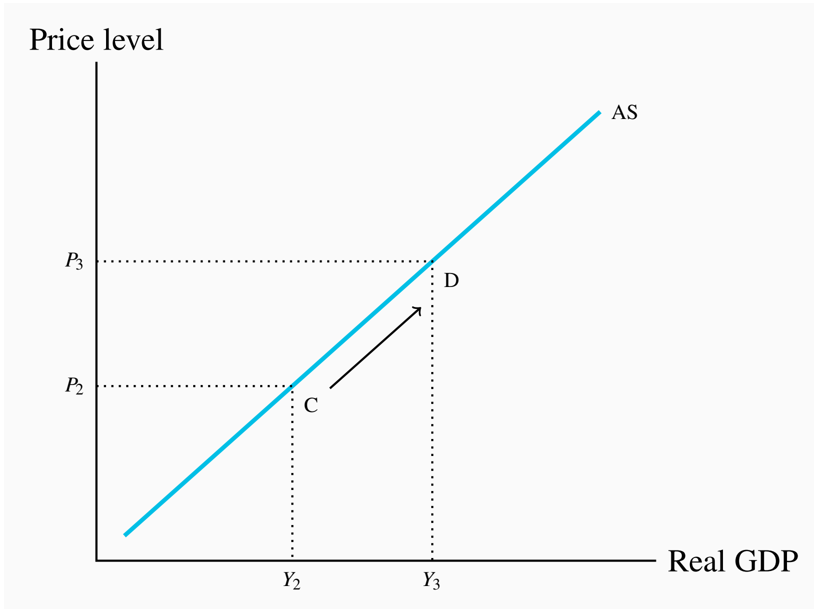
Broadly speaking, in these industries cost per unit of output are increasing with increasing output. Employing more labour and plant capacity means expanding into less productive land and natural resource inputs. Mining gold or extracting bitumen from oil sands are good examples. A rise in price justifies expanding the output of higher cost mines and oil wells. However, many raw material markets are like this including those for agricultural products, forestry products, base metals and natural gas. When market price changes these producers respond by changing their outputs.

In other parts of the economy producers are price setters. Major manufacturing and service industries like auto producers, banks and wireless phone companies face market conditions that are different from those of commodity producers. They set prices based on costs of production and sales and profit targets, and supply the number of cars or bank services or cell phone accounts that are in demand at those prices.

In these industries costs per unit of output are constant over a wide range of current outputs. Money wage rates are fixed, the capacity to produce output is flexible and productivity is constant. If demand for their product or service increases they can supply more by hiring more employees at existing wage rates and selling more output at existing prices. Industries like major manufacturing, retail services, financial services, hospitality services, and professional services are some examples. Output and changes in output are determined by demand.

The upward-sloping aggregate supply curve in Figure 17.3 captures both market conditions to show the output producers are willing to produce and the price level. The aggregate supply curve is drawn based on the assumptions that money wage rates and all other conditions except price that might affect output decisions are constant. As we will see in later chapters, money wage rates and productivity are the most important of these conditions. They determine the *position* of the AS curve.

Figure 17.3: The aggregate supply curve



The AS curve shows the relationship between price level and real GDP, assuming the prices of factors of production are constant. The position of the curve is determined by factor prices and productivity. The slope is determined by changes in costs of production and producer price decisions as output changes.

The *slope* of the AS curve depends on changes in cost per unit of output and price changes if aggregate output changes. As a result it reflects the structure of industry. In Canada, for example Table 16.2 shows that about 70 percent of real GDP comes from service producing industries. Consequently we would expect a smaller positive slope in the AS curve than in Figure 17.3.

In Figure 17.3, if price were P_2 the AS curve shows that business would be willing to produce aggregate output Y_2, which would generate an equal flow of real income. A rise in aggregate output from Y_2 to Y_3 would mean a rise in price to P_3 to meet the increased costs and profits associated with output at this level. Changes in output or price, holding all other conditions constant, move the economy along the AS curve. Moving from point C to point D in the diagram shows this relationship.

On the other hand, a change in any of the conditions assumed to be constant will shift the entire AS curve. A rise in money wage rates (W), for example, would increase labour costs per unit of output (W/Y) at every level of output. The AS curve would shift up vertically as prices rose in order to cover the increased unit labour costs.

17.2 Equilibrium output and potential output

The distinction between equilibrium output and **potential output** is very important to our study of the economy. In the short run, AD and AS determine *equilibrium output*. *Potential output* is determined by the size of the labour force, the stock of capital, and the state of technology. The general level of prices and short-run aggregate demand and supply conditions do not affect potential output.

> **Potential output**: the real GDP the economy can produce on a sustained basis with current labour force, capital and technology without generating inflationary pressure on prices.

Short-run equilibrium real GDP is determined by AD and AS conditions. Fluctuations in real GDP and price are a result of short-run changes in economic conditions. To evaluate the economy's performance and understand how it behaves over time, we need a benchmark. Potential output is the output the economy can produce on a sustained basis using the current labour force, capital, and technology without putting continuous upward pressure on the price level or the inflation rate.

In the short run, the labour force, the capital stock, and technology are fixed by assumption. Potential output is the economy's output based on "full employment" of these inputs, but it is not the maximum output an economy can conceivably make. For short periods of time we could make more by using labour for longer hours and factories for extra production shifts. Just as a marathon runner can sprint from time to time but cannot sustain the sprint over the whole race, the economy can operate for short time periods at levels of output above potential. Potential output is the output the economy can produce on a sustained basis.

When the economy is at potential output, every worker wanting a job at the equilibrium wage rate can find a job, and every machine that can be profitably used at the equilibrium cost for capital is in use. Thus, potential output includes an allowance for "equilibrium unemployment" or structural unemployment and some excess capacity. Some people, who would work at higher wage rates, do not want to work at the equilibrium wage rate. Moreover, in a constantly changing economy, some people are joining the labour force, others are leaving, and still others are temporarily between jobs. Today, Canadian potential output means an unemployment rate of about 6 percent. This is usually called the **natural unemployment rate**.

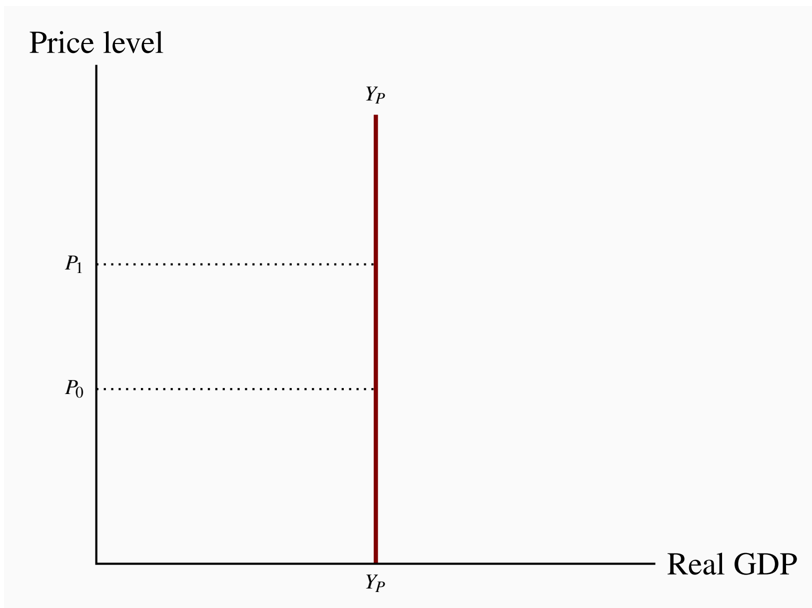
> **Natural unemployment rate**: the unemployment rate that corresponds to potential GDP.

Actual output can also fall below potential output. Workers who want jobs may be unemployed, and producers may have idle plant and equipment or excess capacity. The unemployment rate rises above the 6 percent "full employment" rate.

A key issue in macroeconomics is the way differences between actual output and potential output affect unemployment rates, wage rates, and inflation rates. These effects are important to how the economy provides the standard of living and the way it might adjust equilibrium output to potential output without policy intervention.

Figure 17.4 illustrates potential real GDP (Y_P) with a vertical line. Changes in price from P_0 to P_1, for example, have no effect on Y_P. Changes in the supply of labour, the stock of capital, or the state of technology would increase potential output and shift the vertical Y_P line to the right or to the left.

Figure 17.4: Potential GDP

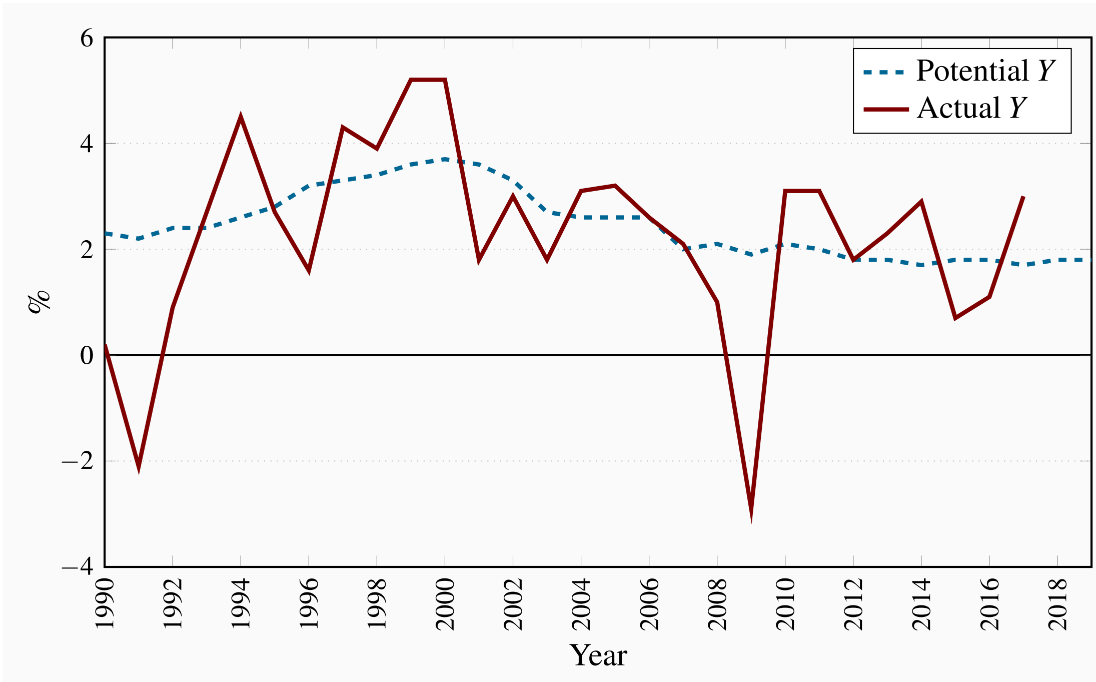


Potential GDP (Y_P) is the real GDP the economy could produce on a sustained basis without putting pressure on costs and prices. Y_P is independent of P.

17.3 Growth in potential output

Growth in the labour force and improvements in labour productivity increase the economy's potential output over time. Labour productivity grows as a result of advances in technology and knowledge coming from investments in capital equipment, education and training. Figure 17.5 shows estimated growth rates for potential and actual real GDP each year in Canada over the period from 1990 to 2018. Potential GDP grew over this period, reflecting the underlying growth in labour force, the stock of capital, and improved technology. But annual growth rates were not constant and in the period since 2000 have tended to decline. Part of this decline is attributed to lower rates of productivity growth in recent years as compared to earlier periods.

Figure 17.5: Annual Growth in Potential and Actual Real GDP in Canada, 1990–2018



Sources: Statistics Canada Table 36-10-0222-01, Office of the Parliamentary Budget Officer, *PBO's Approach to Measuring Potential GDP*, Bank of Canada, *Monetary Policy Reports* and author's calculations.

Figure 17.5 shows how growth rates in actual GDP are more volatile relative to growth rates in potential output. The negative growth rates in 1991 and 2009 mark the recessions of those years. Fluctuations in AD and AS cause business cycles in real GDP and employment. Unemployment rises when output growth is less than the growth in potential output and falls when it is greater.

17.4 Business cycles and output gaps

In some years GDP grows very rapidly, and in other years it actually falls. Growth of potential GDP is also variable but to a lesser degree. These up and down fluctuations in the growth of real GDP are described as **business cycles** in economic activity.

Business cycles: short-term fluctuations of actual real GDP.

Business cycles cause differences between actual and potential GDP. **Output gaps** measure these differences. In a short-run aggregate demand and supply model with a constant potential output, the gap is:

$$\text{Output Gap} = Y - Y_P \tag{17.1}$$

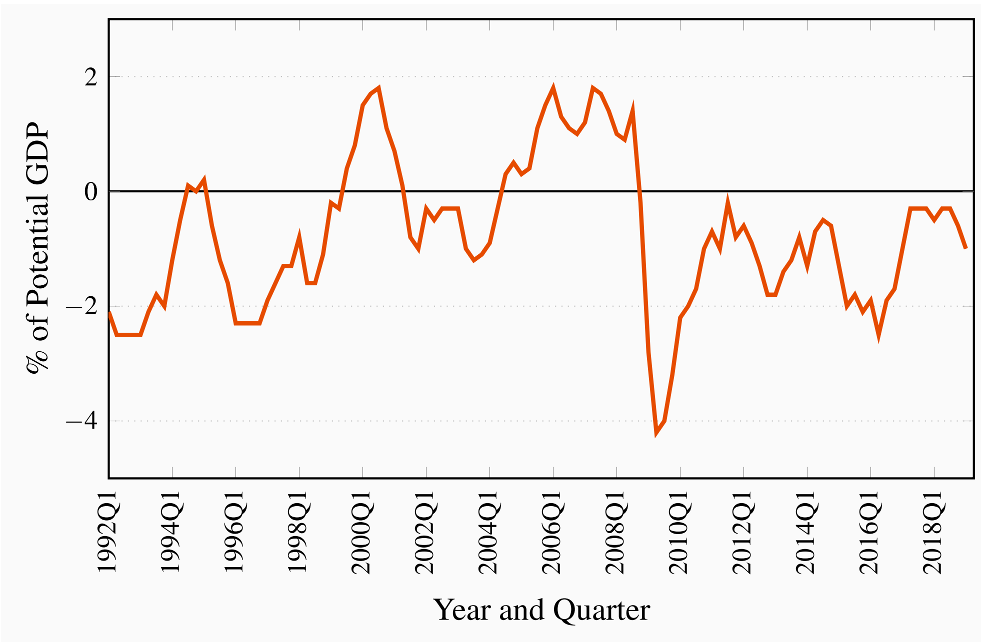
Output gap: the difference between actual output and potential output.

In an economy that grows over time the output gap $Y - Y_P$ is usually measured relative to potential output. This recognizes that a gap of \$10 million is a more serious matter in an economy with a potential output of \$1,000 million than in an economy with a potential output of \$5,000 million.

Figure 17.6 plots the Bank of Canada's estimates of the differences between actual and potential GDP for Canada, quarterly for each year from 1992Q1 to 2019Q1, expressed as a percentage of potential GDP, calculated as:

$$\text{Output Gap}(\%) = \frac{Y - Y_P}{Y_P} \times 100 \tag{17.2}$$

Figure 17.6: The Output Gap in Canada, 1992Q1–2019Q1



Source: Bank of Canada, Indicators of Capacity and Inflation Pressures for Canada, www.bankofcanada.ca.

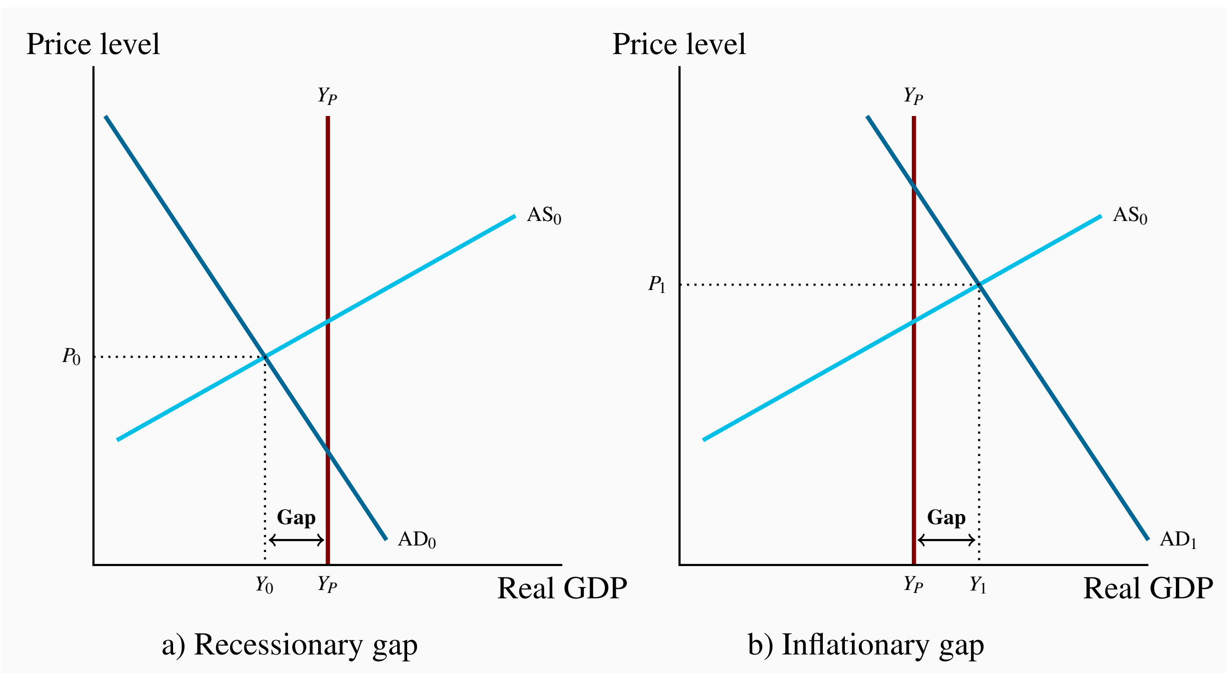
When we compared growth in actual real GDP and potential GDP in Canada from 1990 to 2018 in Figure 17.5, we see an example of the business cycle caused by the 2008 – 2009 financial crisis. Real GDP growth declined in 2008 to less than growth potential GDP and turned negative in 2009. This was a *recession*. It created the negative output gap in Figure 17.6 that starts in the first quarter of 2009 and persists in the most recent data for the first quarter of 2019.

Output gaps describe and measure the short-run economic conditions, and indicate the strength or weakness of the economy's performance. High growth rates in the boom phase of the cycle create positive output gaps, which are called **inflationary gaps** because they put upward pressure on costs and prices. Low or negative growth rates that result in negative output gaps and rising unemployment rate are called **recessionary gaps**. They put downward pressure on costs and prices. As economic conditions change over time, business cycle fluctuations move the economy through recessionary and inflationary gaps. However, you will notice in Figure 17.6 that recessionary gaps in Canada have been deeper and more persistent than inflationary gaps over the past 30 years.

> **Inflationary gap**: a measure of the amount by which actual GDP is greater than potential GDP.
>
> **Recessionary gap**: a measure of the amount by which actual GDP is less than potential GDP.

We can show output gaps in diagrams using the aggregate demand and supply curves and the potential output line introduced in Figures 17.1 and 17.4. Figure 17.7 provides an example. Panel a) illustrates a recessionary gap. Panel b) shows an inflationary gap.

Figure 17.7: Output Gaps



In Figure 17.7 panel a) shows a recessionary gap because AD is too weak to support equilibrium at Y_P. Alternatively in panel b) AD is too strong and pushes equilibrium past Y_P.

The AD and AS model provides a basic explanation of the differences we see between actual real GDP and potential real GDP. Short-run AD and AS conditions determine equilibrium real GDP, which may be either greater or less than potential GDP. Furthermore, the business cycles are the results of changes in the short-run AD and AS conditions.

Introduction to the AD/AS model is an important first step in the study of the performance of the macro economy. But there are more questions:

1. How would the economy react to a persistent output gap?
2. Why do short-run AD and AS conditions change from time to time?

The first question is considered in the remainder of this chapter. Chapter 18 starts on the second question.

17.5 Adjustments to output gaps?

Potential output is real GDP when all markets are in equilibrium. Output gaps indicate disequilibrium in some markets. If we leave the short run and drop the assumption that factor prices are constant, we can ask:

How does the economy react to persistent output gaps?

The answer to this question depends in part on the *flexibility of wage rates and prices* and in part on how *planned expenditure* responds to the flexibility in wage rates and prices.

We also know from national accounts that labour costs are the largest part of factor costs of production, and labour costs per unit of output are the largest part of prices. If the labour market is not in equilibrium—which means unemployment rates not equal to the natural rate—this results in changes in money wage rates. Persistent output gaps will change wage rates and other factor prices and costs. Changes in costs will change prices, shifting the short-run AS curve. The economy may have an adjustment mechanism that tries to eliminate output gaps over time.

This adjustment process assumes that the AD curve is not changed by the fall in wage rates that shifts the AS curve. There are good reasons for skepticism here. In a money economy, with debt and financial contracts denominated in nominal terms, a general fall in money incomes can cause financial distress, extensive insolvencies and reductions in AD. As a result, economists generally agree that *deflation*, a persistent fall in the general *price level*, is *contractionary*, not expansionary as the simple adjustment process suggests. This is reflected in the current concerns that deflation might return to Japan and emerge in Europe as growth stagnates and inflation rates have fallen.

17.6 The role of macroeconomic policy

In Chapter 16, performance of the economy was evaluated based on the standard of living, measured as the real GDP per capita, it provided. Recessionary gaps reduce the standard of living in the economy by reducing employment, real GDP, and per capita real GDP.

Inflationary gaps reduce standards of living in more subtle ways. They push up the price level, raising the cost of living. But the rise in the cost of living affects different people in different ways. Those on fixed money incomes suffer a reduction in their standards of living. People holding their wealth in fixed price financial assets like bank deposits and bonds suffer a loss in their real wealth. On the other hand, inflation reduces the real value of debt, whether it is mortgage debt used to finance the purchase of a house, or a student loan used to finance education. The money repaid in the future has a lower purchasing power than the money borrowed. In these and other ways, the costs of inflation are distributed unevenly in the economy, making decisions about employment, household expenditure, and investment more difficult.

We have also seen, in Figure 17.6, that output gaps have been persistent in the Canadian economy despite the possibility that flexible wages and prices might automatically eliminate gaps. These observations raise two questions:

1. Why are output gaps, especially recessionary gaps, persistent?
2. Can government policy work to eliminate output gaps?

To answer the first question, we need to think about two issues. The first is the *flexibility or rigidity of wages and prices* both up and down. The second is the strong possibility of asymmetry between adjustment effects of absolute increases and absolute decreases in wages and prices. These are topics for later discussion.

The important immediate policy question is: When wages and prices are sticky, should government wait for a self-adjustment process to work, accepting the costs of high unemployment or rising inflation that it produces? This was a very serious and widely debated question since 2008 in the face of growing international recessions as a consequence of serious government debt problems and continued international financial market uncertainty.

Government has policies it can use to reduce or eliminate output gaps. In Chapter 19 we will examine **fiscal policy**, the government expenditures and tax policy that establish the government's budget and its effect on aggregate demand. Government can use its fiscal policy to change the AD curve and eliminate an output gap without waiting for the economy to adjust itself. Chapters 21 and 22 discuss **monetary policy**, actions by the monetary authorities designed to change aggregate demand and eliminate output gaps by changing interest rates, money supply, and the availability of credit. Both fiscal and monetary policy work to change aggregate demand and eliminate output gaps, which reduce the standard of living the national economy provides for its citizens.

> **Fiscal policy**: government expenditure and tax changes designed to influence AD.
>
> **Monetary policy**: changes in interest rates and money supply designed to influence AD.

Application Box 17.1: The Impact of COVID-19 policies in AD/AS

Policies to control the spread of COVID-19 affect real GDP and the price level by changing AS and AD. The decline in AS was indicated by the April 28, 2020 report by Service Canada that applications from 7.3 million Canadians for financial aid had been received. Statistics Canada found one in five Canadian businesses had laid of more than 80 percent of their staff. The fall in AD is indicated by an April 29, 2020 Statistics Canada report that over half of all businesses saw more than a 20 percent decline in revenue.

In early April 2020 the Parliamentary Budget Office reported the following estimates of the impacts on output and inflation.

Table 17.1: Estimates of real GDP and GDP inflation

Indicator:	2019Q4	2020Q1	2020Q2	2020Q3	2020Q4	2020
Real GDP growth (%)	−0.3	−2.5	−25.0	0	5.0	−5.1
Index of GDP deflator (%)	4.1	0.0	−10.0	0	1.0	−0.9

Source: Parliamentary Budget Office: Scenario Analysis Update: COVID-19 Pandemic and Oil Price Shocks, April 9, 2020. Revised PBO Report. Appendix A. https://www.pbo-dpb.gc.ca/web/default/files/Documents/Reports/RP-2021-004-S/RP-2021-004-S_en.pdf

Key Concepts

The **Aggregate demand and supply** model provides a framework for our study of the operation of the economy.

Aggregate demand is the negative relationship between planned aggregate expenditure on final goods and services and the price level, assuming all other conditions in the economy are constant.

Aggregate supply is the positive relationship between outputs of goods and services and the price level, assuming factor prices, capital stock, and technology are constant.

Short-run equilibrium real GDP and price are determined by short-run aggregate demand and aggregate supply, illustrated by the intersection of the AD and AS curves.

Potential output is the output the economy can produce on an ongoing basis with given labour, capital, and technology without putting persistent upward pressure on prices or inflation rates.

The **Natural unemployment rate** is the 'full employment' unemployment rate observed when the economy is in equilibrium at potential output.

Growth in potential output comes from growth in the labour force and growth in labour productivity coming from improvements in technology as a result of investment in fixed and human capital.

Business cycles are the short-run fluctuations in real GDP and employment relative to Potential Output (GDP) and full employment caused by short-run changes in aggregate demand and supply.

Output gaps are the differences between actual real GDP and potential GDP that occur during business cycles.

Unemployment rates fluctuate with output gaps.

Inflationary gaps and recessionary gaps are the terms used to describe positive and negative output gaps based on the effects the gaps have on factor prices.

Actual output adjusts to potential output over time *if factor input and final output prices are flexible* and changes in prices shift the aggregate supply curve to equilibrium with aggregate demand at Y_P.

Fiscal and monetary policy are tools governments and monetary authorities can use to stabilize real output and employment or speed up the economy's adjustment to output gaps.

Exercises for Chapter 17

Exercise 17.1 Suppose we have the following information for an economy:

GDP deflator	Planned aggregate expenditure	Planned aggregate output
90	550	150
100	500	300
110	450	450
120	400	600
130	350	750

(a) Plot the AD and AS curves in a carefully labeled diagram.

(b) What are the short-run equilibrium values of real GDP and the price level?

Exercise 17.2 Suppose we learn that potential output is 500 for the economy in Exercise 17.1.

(a) Add a line to your diagram for Exercise 17.1 to illustrate potential GDP.

(b) What is the size of any output gap you see in the diagram?

Exercise 17.3 Potential GDP is determined by the size of the labour force, the stock of capital and the state of technology used in the production process. Assume the labour force grows over time, and research and development lead to improvements in technology, and productivity. Use an AD/AS diagram to illustrate potential GDP both before and after the growth in labour force and the improvement in technology.

Exercise 17.4 Growth in potential output is determined by growth in the labour force and growth in labour productivity. Suppose the labour force grows by 1.5 percent a year and labour productivity, based on increased capital and improved technology, grows by 1.0 percent a year.

(a) What is the annual growth in potential output?

(b) Illustrate the growth in potential output in an AD/AS diagram.

(c) Aggregate demand is not changed by the change in potential output. Indicate any output gap caused by the change in potential output.

Exercise 17.5 Suppose we have the following data for an economy:

Year	Potential output (billions 2007$)	Real GDP (billions 2007$)
2006	1,038	1,017
2007	1,069	1,030
2008	1,101	1,101
2009	1,134	1,160
2010	1,168	1,139
2011	1,203	1,130
2012	1,240	1,187
2013	1,277	1,163

Calculate the output gap for each year in this economy. Plot the output gap in a time series diagram. Date the timing of the phases of any business cycles you see in your plot of the output gap.

Exercise 17.6 Draw an AD/AS diagram that shows an economy called Westland in short-run equilibrium with GDP equal to potential GDP.

(a) Suppose a slowdown in the rate of growth of GDP in China cuts Chinese imports of primary products from Westland. Using your AD/AS diagram illustrate and explain your forecast for the effects of this change on the equilibrium GDP and price level in Westland.

(b) What effect, if any, would a slowdown in GDP growth in China have on employment and unemployment rates in Westland? Explain why.

Exercise 17.7 *Optional*: Consider an economy described by the following: AD: $Y = 2250 - 10P$, AS: $P = 125 + 0.1Y$.

(a) What are the short-run equilibrium values for real GDP and the price level?

(b) Assume potential output is 500 and draw an AD/AS/Y_P diagram to show the initial short-run equilibrium real GDP, price level and potential output.

(c) Changes in international market conditions drive up prices for crude oil and base metals. Increased production costs driven by these higher input prices raise the general price level by 5 at every level of output. Write the equation of the new AS curve. What are the new short-run equilibrium real GDP and price level?

(d) Draw the new AS curve in your diagram for (b). What is the size of the output gap?

Chapter 18

Aggregate expenditure & aggregate demand

This chapter explains a basic model of the economy:

The Canadian economy was very much in the news in late 2018 and the first half of 2019. Economic growth in the last quarter of the 2018 had slowed, lowering the annual rate of growth in real GDP from 3.5 percent in 2017 to 2.2 percent in 2018. Early forecasts for growth the year 2019 were mixed but pessimistic on balance. Some commentators said the economy was heading into a 'recession' as defined by two consecutive quarters of negative growth. Business investment expenditures had declined, as had oil prices, and international trade arrangements were uncertain. Indeed the Bank of Canada's estimate of the output gap had changed from -0.3 in 2018Q3 to -1.0 in 2019Q1.

Other forecasts were more optimistic. A recovery in crude oil prices and a rise in exports in early 2019 would offset some of the weakness in demand and income. Unexpectedly strong growth in employment in the first half of 2019 was further grounds for continued growth.

Furthermore there was a change in the monetary policy stance of the Bank of Canada. Monetary policy had played a role in the slow down in growth rates. The Bank of Canada had been raising its overnight interest rate setting, in steps, from 1.00 percent in December 2017 to 1.75 percent by October 2018 to moderate strong growth and inflation pressure. But by December 2018 there was recognition that further increases in interest rates should be postponed. A period of interest rate stability followed with the Bank's policy rate unchanged in scheduled announcements up to and including January 2020.

COVID-19 changed all that. The government mandated business closings and physical distancing rules to contain the pandemic. Households and businesses suffered corresponding losses of incomes and revenues. Short run aggregate demand was cut and experienced continued downward pressure. The Bank of Canada changed its monetary policy to provide financial support. The Bank cut its overnight interest rate in steps from 1.75 percent to 1.25 percent on March 4, 2020, and then to 0.25 percent on March 27. As of April 7th estimated federal government announced fiscal support for the economy totaled $105.5 billion with more to come as needed. On April 9, 2020, Statistics Canada reported that over 3 million Canadians were affected by job losses or reduced hours. Both aggregate demand and aggregate supply were sharply reduced. Even with monetary and fiscal policy support, estimated losses of real GDP in the second quarter of 2020 were in the 15 - 30 percent range. (See: Statistics Canada: https://www150.statcan.gc.ca/n1/daily-quotidien/200429/dq200429a-eng.htm)

Forecasting the effects of changes in economic conditions or government policy shifts is complicated by the complex system of interdependencies and feedback effects in the macro economy.

This chapter introduces a basic short-run model of the economy as a first step in explaining changes in aggregate demand that can change economic performance.

18.1 Short-run aggregate demand and output

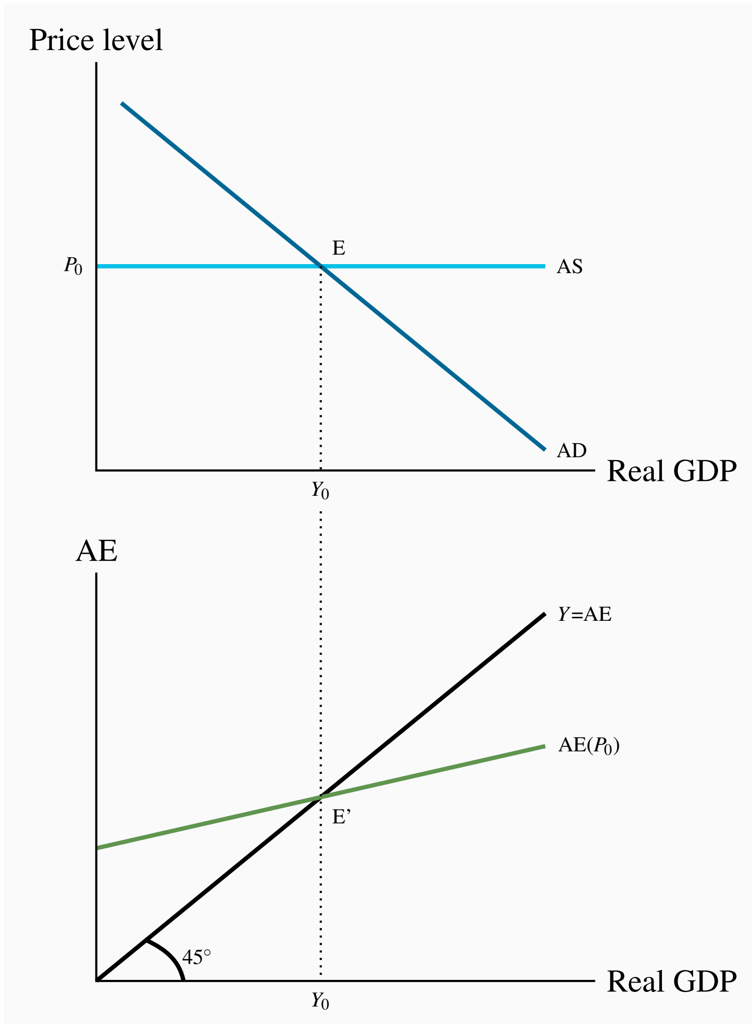
Consider first just the private market sector. Assume there are households and businesses in this simple economy, but no government and no financial markets. The households and businesses buy domestically produced and imported goods and services. Businesses also sell some output in export markets to residents of other countries. This basic model offers a simple but useful example of how the actual economy works.

This initial *short-run* model is based on the following assumptions:

- All prices and wages are fixed at a given level.
- At these prices and wages, businesses produce the output that is demanded and labour accepts opportunities to work.
- Money supply, interest rates and foreign exchange rates are fixed because at this stage we ignore the financial sector.

With constant prices **aggregate demand** determines total output, real GDP. Figure 18.1 uses an AD/AS diagram like those developed in Chapter 17 to illustrate these conditions and develop *an aggregate expenditure* function.

Figure 18.1: Aggregate demand, aggregate expenditure and output when the price level is constant



Aggregate demand determines total output, real GDP.

The horizontal AS curve in the upper part of Figure 18.1 shows that the price level is fixed at P_0. As a result, the equilibrium real GDP in this example is determined by the *position of the AD curve*. Changes in the position of the AD curve would cause changes in real output and real income, and corresponding changes in employment.

The position of the AD curve in the diagram is determined by things, *other than price*, that affect expenditure decisions. Understanding these expenditure decisions and their effects are the focus of this and the next several chapters.

The lower part of Figure 18.1 shows the relationship between **planned aggregate expenditure** (AE) and *income as measured by real GDP(Y)* when the price level is fixed. It also shows that, if the aggregate expenditure function (AE) has the right position and slope, there is a level of output at which planned expenditure and output are equal ($Y = \text{AE}$). Then revenue the business sector receives from sales of current output just covers the costs of production including expected profit. Planned expenditure and planned output are in equilibrium.

> **Aggregate demand** is determined by equality between **planned aggregate expenditure (AE)** and real GDP.

This equality between planned expenditure and output determines the position of the AD curve, as shown in the upper part of the diagram.

The interactions of expenditure, output, and income shown in the lower part of Figure 18.1 define a basic macroeconomic model.

18.2 Aggregate expenditure

The model starts with the expenditure categories defined and measured in national accounts and described in Chapter 16. By the expenditure approach, GDP (Y) is the sum of consumption (C), investment (I), and exports (X), minus imports (IM). Then we can write:

$$\text{GDP}(Y) = C + I + X - IM \tag{18.1}$$

Expenditure as measured by national accounts is the sum of actual expenditures by business and households.

> **GDP(Y)** is the national accounts measure of the sum of *actual expenditure in the economy*.

Aggregate expenditure (AE) is planned expenditure by business and households. The distinction between planned and actual expenditures is a key factor in explaining how the national income and employment are determined.

> **Aggregate expenditure (AE)** is *planned expenditure* by business and households.

Induced expenditure

A simple short-run model of the economy builds on two key aspects of planned expenditure, namely *induced* expenditure and *autonomous* expenditure. First, an important part of the expenditure in the economy is directly related to GDP and changes when GDP changes. This is defined as **induced expenditure**. It is mainly a result of household expenditure plans and expenditure on imported goods and services embodied in those plans.

> **Induced expenditure** is planned expenditure that is determined by current income and changes when income changes.

The largest part of consumption expenditure by households is *induced expenditure*, closely linked to current income. This expenditure changes when income changes, changing in the same direction as income changes but changing by less than income changes. The **marginal propensity to consume** ($mpc = c$) defines this link between changes in income and the changes in consumption they induce. The mpc is a positive fraction.

> **Marginal propensity to consume** ($mpc = c = \Delta C/\Delta Y$) is the change in consumption expenditure *caused by* a change in income.

Household's expenditures on imports are at least partly induced expenditures that change as income and consumption expenditure changes. The **marginal propensity to import** ($mpm = m$) defines this link between changes in income and the changes in imports. The mpm is also a positive fraction because changes in income induce changes in imports in the same direction but by a smaller amount.

> **Marginal propensity to import** ($mpm = m = \Delta IM/\Delta Y$) is the change in imports caused by a change in income.

Because expenditure on imports is not demand for domestic output, aggregate **induced expenditure** for the economy is determined by *subtracting* induced import expenditure from induced consumption expenditure to get induced expenditure on domestic output.

> **Induced expenditure** $(c - m)Y$ is planned consumption expenditures net of imports that changes when income changes.

The relationship between expenditures and national income (GDP) is illustrated in Table 18.1 using a simple numerical example and then in Figure 18.2 using a diagram.

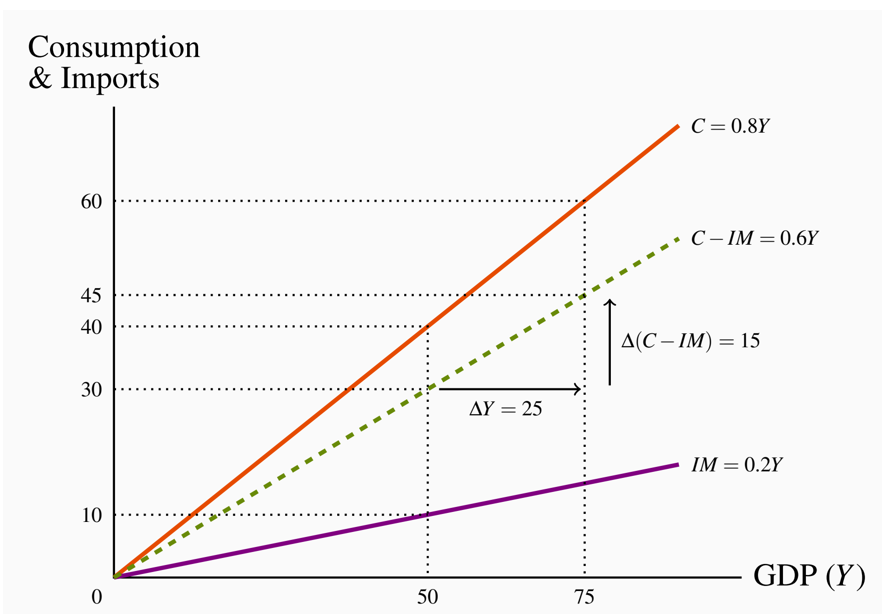
Table 18.1: The effects of changes in GDP on consumption, imports and expenditure.

Consumption function:	$C = 0.8Y$		
Imports function:	$IM = 0.2Y$		
Y	Induced $C = \Delta C/\Delta Y$	Induced $IM = \Delta IM/\Delta Y$	Induced expenditure $(\Delta C - \Delta IM)/\Delta Y$
0	0	0	0
50	40	10	30
100	80	20	60
75	60	15	45

The first row of the table illustrates clearly that if there were no income there would be no induced expenditure. The next two rows show that positive and increased incomes cause (induce) increased expenditure. The final row shows that a fall in income reduces expenditure. The induced expenditure relationship causes both increases and decrease in expenditure as income increases or decreases.

Figure 18.2 shows the same induced expenditure relationships in a diagram.

Figure 18.2: Induced expenditures



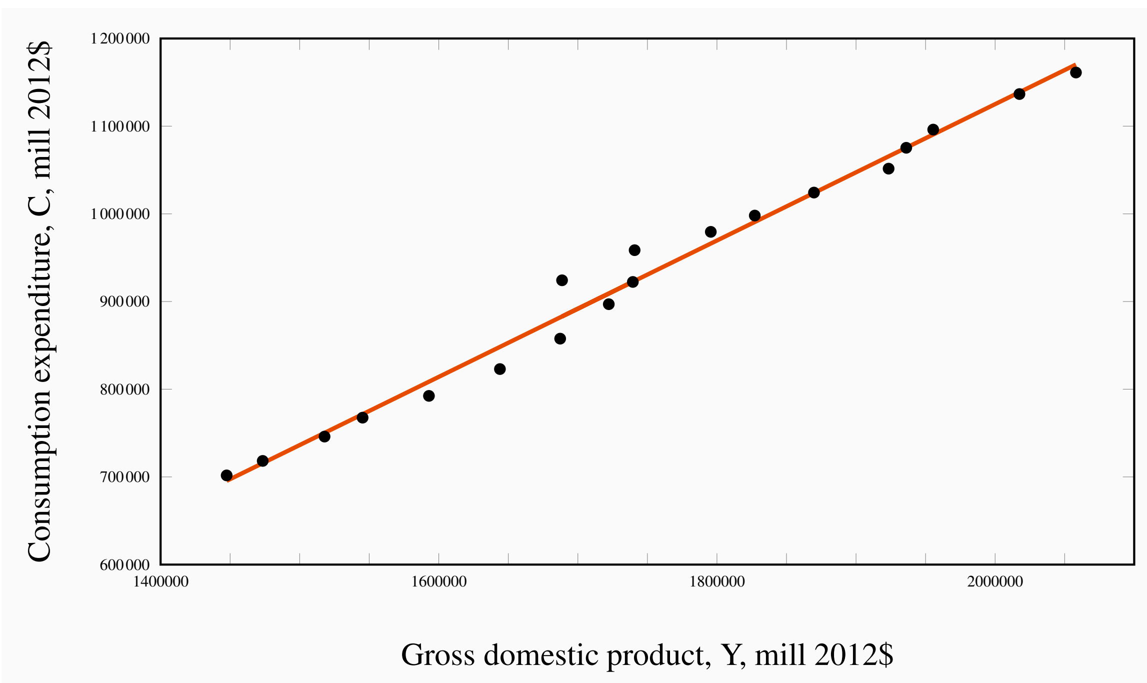
The slope of the $C - IM$ line is: $\Delta(C - M)/\Delta Y = 15/25 = 3/5 = 0.6$.

Positive slopes of the consumption and import lines in the diagram show the positive relationship between income and expenditure. A rise in income from 0 to 50 induces a rise in aggregate expenditure from 0 to 30 because consumption expenditure increases by 40 of which 10 is expenditure on imports. If income were to decline induced expenditure would decline, illustrated by movements to the left and down the expenditure lines.

The marginal propensities to spend are the slopes of the expenditure lines. If, for example, GDP increased from 50 to 75 in the diagram, consumption expenditure on domestic goods and services $C - IM$ would increase by 15. The slope of the $C - IM$ line is defined as rise/run which is $15/25 = 0.6$. By similar observations and calculations the slopes of the C line and the IM line are $\Delta C/\Delta Y = 0.8$ and $\Delta IM/\Delta Y = 0.2$.

These induced expenditure relationships are fundamentals of a basic macro model. The numerical values used here are chosen just to illustrate the relationships. Values used in the model of an actual economy, like the Canadian economy, would be estimated by econometric analysis of Canadian data. It is important that these relationships remain stable even as economic conditions change. The close statistical relationship between income and consumption in Figure 18.3 and size and stability of consumption share in expenditure GDP are strong evidence in support of that stability.

Figure 18.3: A Consumption Function for Canada, 2000–2018



Source: Statistics Canada Table 36-10-0222-01 Gross Domestic Product, expenditure-based, (x 1,000,000 2012$)

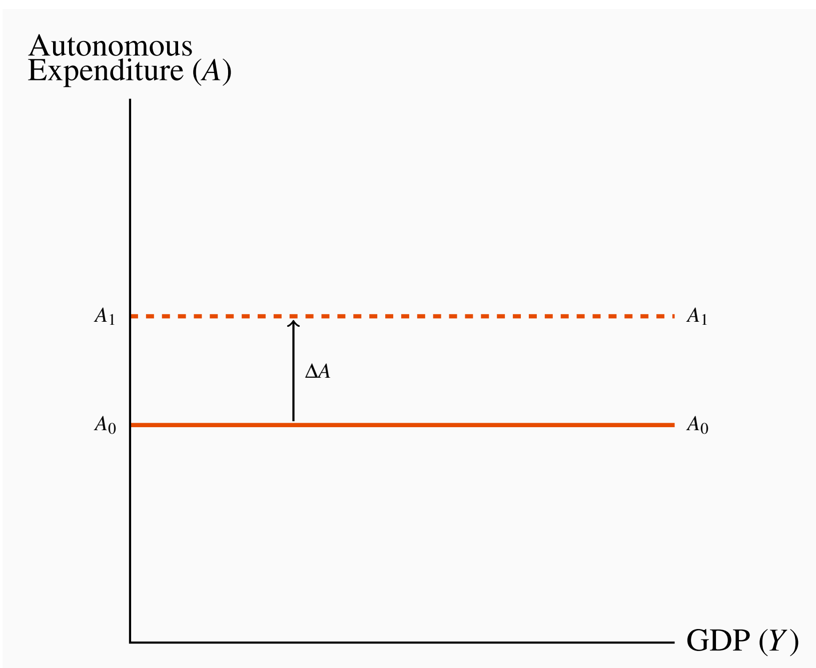
Autonomous expenditure

The second part of GDP expenditure is where the real action lies. It covers the important changes in expenditure that drive the business cycles in economic activity. **Autonomous expenditure** (A), is the planned expenditure that is not determined by current income. It is determined instead by a wide range of economic, financial, external and psychological conditions that affect household and business decisions to make expenditures on current output.

> **Autonomous expenditure** (A) is planned expenditure that is not determined by current income.

Consumption and imports have an autonomous component in addition to the induced component. But investment (I) and exports (X) are the major autonomous expenditures. Figure 18.4 shows the independence of autonomous expenditure from income. Unlike induced expenditure, autonomous expenditure does not change as a result of changes in GDP. The slope of the A line in the diagram is zero ($\Delta A/\Delta Y = 0$). However, a change in autonomous expenditure would cause a parallel shift in the A line either up or down. An increase in A is shown.

Figure 18.4: Autonomous expenditure



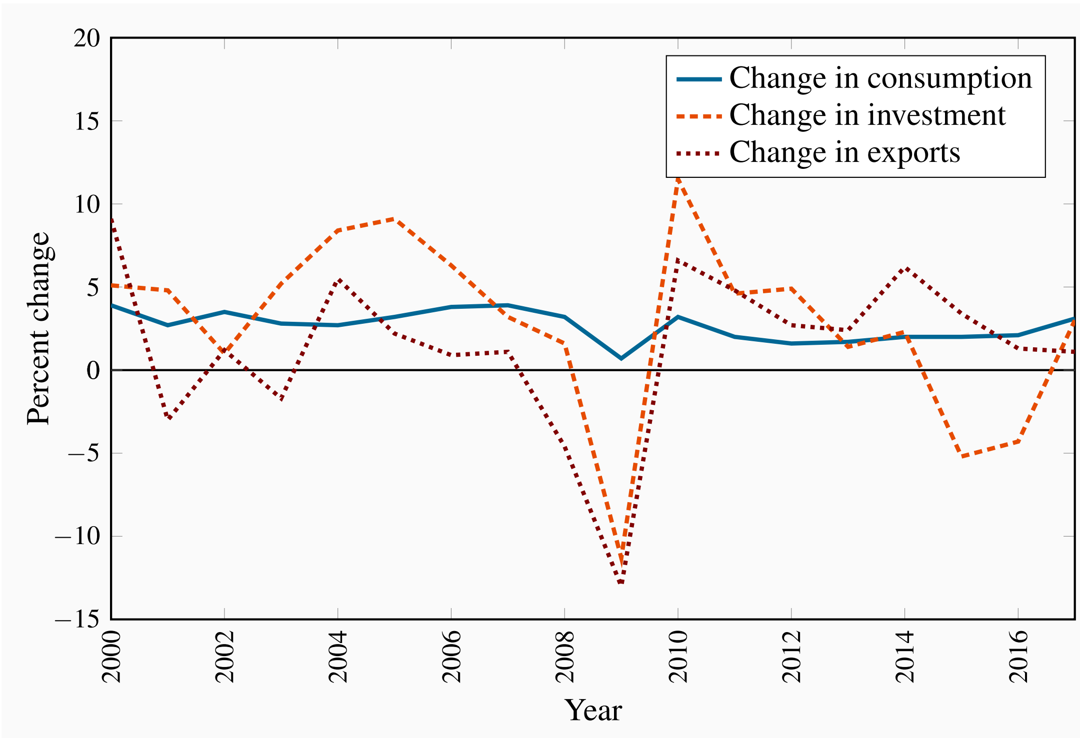
Investment expenditure (I) is one volatile part of aggregate expenditure. It is expenditure by business intended to change the fixed capital stock, buildings, machinery, equipment and inventories they use to produce goods and services. In 2018 investment expenditures were about 27 percent of GDP. Business capacity to produce goods and services depends on the numbers and sizes of factories and machinery they operate and the technology embodied in that capital. Inventories of raw materials, component inputs, and final goods for sale allow firms to maintain a steady flow of output and supply of goods to customers.

Firms' investment expenditure on fixed capital depends chiefly on their current expectations about how fast the demand for their output will increase and the expected profitability of higher future output. Sometimes output is high and rising; sometimes it is high and falling. Business expectations about changes in demand for their output depend on many factors that are not clearly linked to current income. As a result we treat investment expenditure as *autonomous*, independent of current income but potentially volatile as financial conditions and economic forecasts fluctuate. For example, the sharp drop in crude oil prices in late 2017 and again in the spring of 2020 changed market and profit expectations dramatically for petroleum producers and for a wide range of suppliers to the industry. Investment and exploration projects were cut back sharply, reducing investment expenditure.

Exports (X), like investment, can be a volatile component of aggregate expenditure. Changes in economic conditions in other countries, changes in tastes and preferences across countries, changes in trade policies, and the emergence of new national competitors in world markets all impact on the demand for domestic exports.

To illustrate this volatility in exports and in investment, Figure 18.5 shows the year-to-year changes in investment, exports, and consumption expenditures in Canada from 1999 to 2017. You can see how changes in investment and exports were much larger than those in consumption. This volatility in investment and exports causes short-term shifts in aggregate expenditure which cause business cycle fluctuations in GDP and employment. More specifically, the data show the sharp declines in investment and exports that caused the Great Recession in 2009 and even though these autonomous expenditures increased in 2010 those increases were not large or persistent enough to restore pre-recession levels of autonomous expenditure. Indeed investment expenditures declined again in 2015 and 2016 before recovering in 2017 and 2018.

Figure 18.5: Annual percent change in Real Consumption, Investment and Exports Canada 2000–2017



Source: Statistics Canada Table 36-10-0222-01

The aggregate expenditure function

The **aggregate expenditure function (AE)** is the sum of *planned* induced expenditure and *planned* autonomous expenditure. The emphasis on 'planned' expenditure is important. Aggregate expenditure is the expenditure households and businesses want to make based on current income and expectations of future economic conditions. Expenditure based GDP in national accounts measures, after the fact, the expenditures that were made. These may not be the same as planned expenditure. Some plans may not work out.

> **Aggregate expenditure (AE)** is the sum of *planned induced* and *autonomous expenditure* in the economy.

Aggregate expenditure (AE) equals the sum of a specific level of autonomous expenditure (A_0) and induced expenditure $[(c-m)Y]$ or in the simple notation introduced above:

$$\text{AE} = A_0 + (c-m)Y \tag{18.2}$$

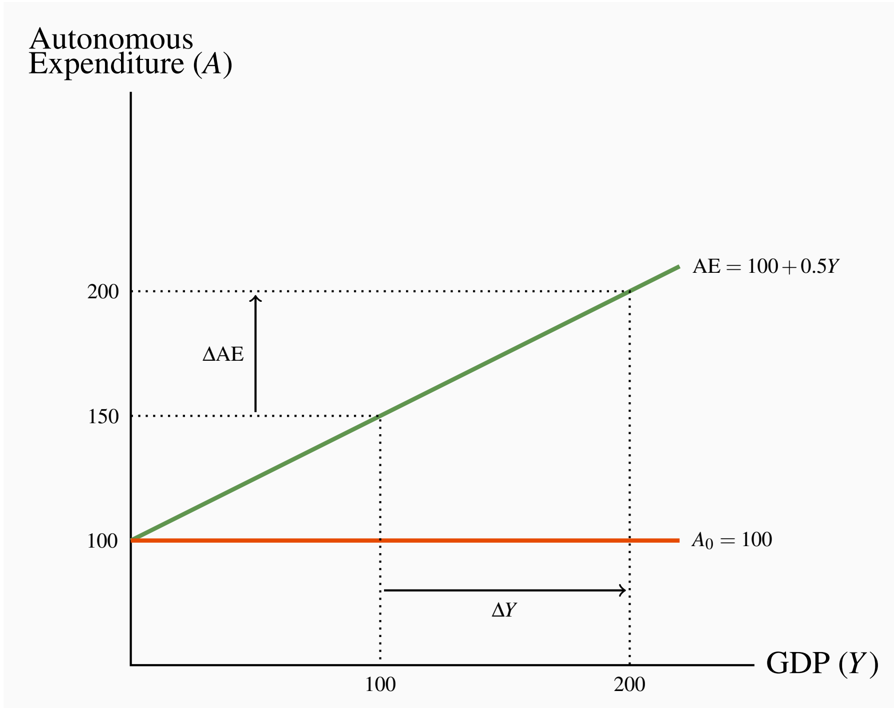
Table 18.2 gives a numerical example of the aggregate expenditure function, assuming the marginal propensity to spend on domestic output $(c-m) = 0.5$. It shows constant autonomous expenditure at each level of GDP, induced expenditure changing as GDP changes and aggregate expenditure as the sum of autonomous and induced expenditure at each level of GDP. Induced expenditure changes in the same direction as GDP which raises aggregate expenditure in rows 3 to 5 in the table but lowers aggregate expenditure in the last row when GDP falls.

Table 18.2: An aggregate expenditure function.

GDP (Y)	**Autonomous Expenditure** $(A_0 = 100)$	**Induced Expenditure** $(c-m)Y = 0.5Y$	**Aggregate Expenditure** **AE**$= 100 + 0.5Y$
0	100	0	100
50	100	25	125
100	100	50	150
175	100	87.5	187.5
200	100	100	200
150	100	75	175

Figure 18.6 shows this aggregate expenditure function with a positive intercept on the vertical axis and positive slope. The positive intercept measures autonomous expenditure. The slope measures induced expenditure. Changes in autonomous expenditure would shift the AE function vertically, up for an increase or down for a decrease. Induced expenditures are usually assumed to be stable in the short run but if they were to change, as a result of change in marginal propensity to import for example, the slope of the AE function would change.

Figure 18.6: An aggregate expenditure function



The AE function has a vertical intercept equal to autonomous expenditure 100 and a slope of $\Delta\text{AE}/\Delta Y = 0.5$ equal to the changes in expenditure induced by changes in Y.

18.3 Aggregate expenditure and equilibrium output

National Accounts as in Chapter 16 measure actual expenditure, output and national income and GDP(Y). The Aggregate Expenditure function gives *planned* expenditure (AE). In a modern industrial economy actual output and income may differ from what was planned, either on the output side or on the purchase and sales side. A simple example of the time sequence of output and sales shows why.

In most cases business install capacity and produce output in anticipation of sales in the near future. This is apparent from the stocks of goods offered in most retail outlets or online. Auto manufacturers, for example like to have an inventory of 30 to 60 days of finished vehicle sales available for retail buyers. Coffee and donut shops have coffee and snacks available to customers when they walk in. Even service industries like cell phone companies try to have staff and product on hand and ready to serve customers on demand. In these and other cases producers incur costs that they expect to recover from later sales. If sales don't match expectations some inventories of products build up or fall short, capacity is not matched to demand and some sales opportunities are lost or costs are not recovered.

As a result, if planned expenditure (AE) and GDP are different then plans in some part of the economy have not been realized and there is an incentive to change output. However, when actual output (GDP) is equal to planned expenditure (AE) expenditure and output plans are successful and, unless underlying conditions change there is no incentive to change output.

Equilibrium output

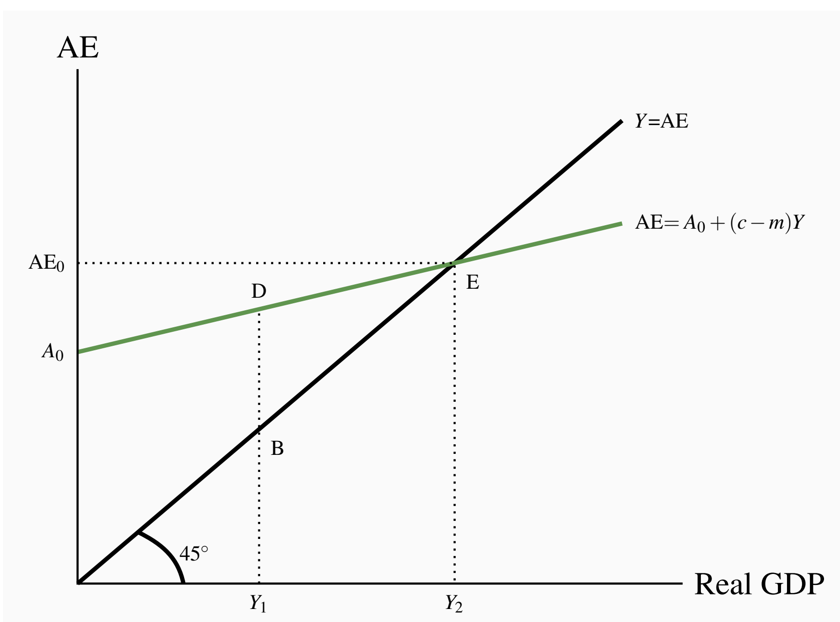
Output is said to be in **short-run equilibrium** when the current output of goods and services equals planned aggregate expenditure:

$$Y = A_0 + (c - m)Y \, (18.2)$$

Then spending plans are not frustrated by a shortage of goods and services. Nor do business firms make more output than they expect to sell. In short-run equilibrium, output equals the total of goods and services households, businesses, and residents of other countries want to buy. Real GDP is determined by aggregate expenditure.

> **Short-run equilibrium output**: Aggregate expenditure and current output are equal ($Y = \text{AE}$).

Figure 18.7: The 45° diagram and equilibrium GDP



The 45° line gives $Y = \text{AE}$ the equilibrium condition. At point E the AE line crosses the 45° line and $\text{AE}_0 = Y_2$. This is the equilibrium. At Y_1, AE$> Y$ and unplanned reductions in inventories provide the incentive to increase Y.

In Figure 18.7 real GDP is measured on the horizontal axis and aggregate spending on the vertical axis. The 45° line labeled $Y = \text{AE}$, illustrates the equilibrium condition. At every point on the 45° line, AE measured on the vertical axis equals current output, Y, measured on the horizontal axis. This 45° line has a slope of 1.

The AE function in Figure 18.7 starts from a positive intercept on the vertical axis to show autonomous aggregate expenditure A_0, and has a slope equal to $(c - m)$ which is less than one. The 45° line has a slope equal to one. At every point on the 45° line, the value of output (and income) measured on the horizontal axis equals the value of expenditure on the vertical axis. With a positive vertical intercept and slope less than 1 the AE line crosses the 45° line at E. Since E is the only point on the AE line also on the 45° line, it is the only point at which output and planned expenditure are equal. *It is the equilibrium point.*

Income Y_2 is the only income at which aggregate expenditure just buys all current output. For example, assume as shown in the diagram, that output and incomes are only Y_1. Aggregate expenditure at D is not equal to output as measured at B. Planned expenditure is greater than current output. Aggregate spending plans cannot all be fulfilled at this current output level. Consumption and export plans will be realized only if business fails to meet its investment plans as a result of an unplanned fall in inventories of goods.

In Figure 18.7 all outputs less than the equilibrium output Y_2, are too low to satisfy planned aggregate expenditure. The AE line is above the 45° line along which expenditure and output are equal. Conversely, if real GDP is greater than Y_2 aggregate expenditure is not high enough to buy all current output produced. Businesses have unwanted and unplanned increases in inventories of unsold goods.

Table 18.3 extends the numerical example in Table 18.2 to show equilibrium when GDP(Y) is 200 and the unwanted inventory changes at other income levels. When GDP in column (1) is less than AE in column (4) current output does not cover current planned expenditure. Inventories fall and producers can't meet their inventory targets. The unwanted change in inventories in column (5) is negative.

Table 18.3: Equilibrium GDP: $Y = \text{AE}$.

GDP (Y) (1)	Autonomous Expenditure ($A_0 = 100$) (2)	Induced Expenditure $(c-m)Y = 0.5Y$ (3)	Aggregate Expenditure $\text{AE} = 100 + 0.5Y$ (4)	Unplanned Δ Inventory $(Y - \text{AE})$ (5)
0	100	0	100	-100
50	100	25	125	-75
100	100	50	150	-50
175	100	87.5	187.5	-12.5
200	100	100	200	0
250	100	125	225	+25
300	100	150	250	+50

You can construct a diagram like Figure 18.7 using the numerical values for GDP and aggregate expenditure in Table 18.3 and a 45° line to show equilibrium $Y_e = 200$. Example Box 18.1 at the end of the chapter illustrates equilibrium for this basic model using simple algebra.

Adjustment towards equilibrium

Unplanned changes in business inventories motivate the adjustments in output that move the economy toward equilibrium output. Suppose in Figure 18.7 the economy begins with an output Y_1, below equilibrium output Y_e. Aggregate expenditure is greater than output Y_1. If firms have inventories from previous production, they can sell more than they have produced by running down inventories for a while. Note that this fall in inventories is unplanned. Planned changes in inventories are already included in planned investment and aggregate expenditure.

> **Unplanned changes in business inventories**: indicators of disequilibrium between planned and actual expenditures – incentives for businesses to adjust levels of employment and output (Y).

If firms cannot meet planned aggregate expenditure by unplanned inventory reductions, they must turn away customers. Either response—unplanned inventory reductions or turning away customers—is a signal to firms that aggregate expenditure is greater than current output, markets are strong, and output and sales can be increased profitably. Hence, at any output below Y_e, aggregate expenditure exceeds output and firms get signals from unwanted inventory reductions to raise output.

Conversely, if output is initially above the equilibrium level, Figure 18.7 shows that output will exceed aggregate expenditure. Producers cannot sell all their current output. Unplanned and unwanted additions to inventories result, and firms respond by cutting output. In general terms, when the economy is producing more than current aggregate expenditure, unwanted inventories build up and output is cut back.

Hence, when output is below the equilibrium level, firms raise output. When output is above the equilibrium level, firms reduce output. At the equilibrium output Y_e, firms sell their current output and there are no unplanned changes to their inventories. Firms have no incentive to change output.

Equilibrium output and employment

In the examples of short-run equilibrium we have discussed, output is at Y_e with output equal to planned expenditure. Firms sell all they produce, and households and firms buy all they plan to buy. But it is important to note that nothing guarantees that equilibrium output Y_e is the level of potential output Y_P. When wages and prices are fixed, the economy can end up at a short-run equilibrium below potential output with no forces present to move output to potential output. Furthermore, we know that, when output is below potential output, employment is less than full and the unemployment rate is higher than the natural rate. The economy is in recession and by our current assumptions neither price flexibility nor government policy action can affect these conditions.

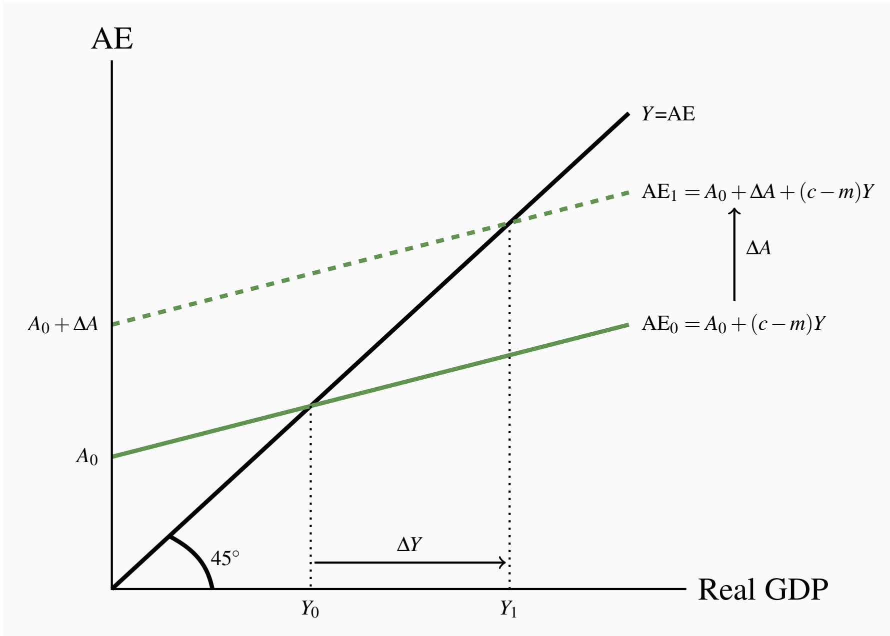
18.4 The multiplier

In our model the slope of the AE line depends on the marginal propensity to consume and the marginal propensity to import. For any given *mpc* and *mpm*, the level of autonomous expenditure (A_0) determines *the height of the AE* line. Recall that autonomous expenditure is expenditure that is not related to national income.

In a diagram, changes in autonomous expenditure cause parallel vertical shifts in the AE function. Autonomous expenditure depends chiefly on current expectations about future domestic and foreign demand for output, future prices and future profits. These expectations about the size and strength of future markets can fluctuate significantly, influenced by current pessimism or optimism about the future. We saw this volatility in investment in Figure 18.6. Similarly, changes in conditions in export markets change exports and changes in consumer confidence change autonomous consumption expenditure.

Suppose firms become very optimistic about future demand for their output. They want to expand their factories and add new equipment to meet this future demand. Autonomous expenditure rises. If other components of aggregate expenditure are unaffected, AE will be higher at each income than before. Figure 18.8 shows this upward shift in AE to AE_1. Before we go into detail, think about what is likely to happen to output. It will rise, but by how much?

Figure 18.8: The effect of a rise in autonomous expenditure



A rise in autonomous expenditure ΔA shifts AE up to AE_1. Equilibrium GDP rises by a larger amount from Y_0 to Y_1.

When autonomous expenditure rises, firms increase output to meet higher demand, increasing their payments for factor inputs to production. Households have higher income and increase their consumption expenditure ($c\Delta Y$) and imports ($m\Delta Y$). Firms increase output again to meet this increased demand, further increasing household incomes. Consumption and imports rise further. This is a first and important example of interdependency and feedback in a basic model. A change in autonomous expenditure, either positive or negative, changes income which in turn causes a change in induced expenditure. Equilibrium income changes by the change in autonomous expenditure plus the change in induced expenditure.

Figure 18.8 shows that an upward shift in the AE function increases equilibrium income by a finite amount, but by a larger amount than the vertical rise in the AE line. This is because $(c-m)$, the slope of AE, is less than unity, giving the AE line a lower slope than the 45° line. Households increase their expenditure when incomes rise, but they increase expenditure by less than the rise in income. Equilibrium moves from Y_0 to Y_1. Equilibrium output rises more than the original rise in investment, $\Delta Y_e > \Delta A$, but does not rise without limit.

A fall in autonomous expenditure would have the opposite effect. AE would shift down and equilibrium income would decline by more than the fall in A.

The **multiplier** is a concept used to define the change in equilibrium output and income *caused* by a change in autonomous expenditure. If A is autonomous expenditure:

$$\text{The multiplier} = \frac{\Delta Y}{\Delta A} \tag{18.3}$$

Multiplier $(\Delta Y/\Delta A)$: the ratio of the change in equilibrium income Y to the change in autonomous expenditure A that caused it.

We can also show the change in equilibrium output caused by a rise in autonomous investment expenditure using the simple numerical example we used earlier. In Table 18.4 initial autonomous expenditure is 100, induced expenditure is $0.5Y$ and initial equilibrium is $Y = 200$. Then autonomous expenditure increases, as in column (3) by 25 to a new level of 125. Induced expenditure is still $0.5Y$. Aggregate expenditure in column (6) rises as a result of increase in both autonomous and induced expenditure.

Table 18.4: The effect of a rise in autonomous expenditure on equilibrium GDP

GDP (Y) (1)	Initial Autonomous Expenditure ($A_0 = 100$) (2)	New Autonomous Expenditure ($A_1 = 125$) (3)	Induced Expenditure $(c-m)Y = 0.5Y$ (4)	Initial Aggregate Expenditure ($\text{AE} = 100 + 0.5Y$) (5)	New Aggregate Expenditure ($\text{AE}_1 = 120 + 0.5Y$) (6)
175	100	-	87.5	187.5	187.5
200	100	-	100	200	200
225		125	112.5		237.5
237.5		125	118.75		243.5
243.5		125	121.75		246.75
246.75		125			
–		125			
250		125	125		250
300		125	150		275

The initial increase in aggregate expenditure illustrated is from 200 to 237.5 made up of an increase in autonomous expenditure of 25 and in induced expenditure of $0.5 \times 25 = 12.5$. However GDP(Y) is only 225 and planned expenditure is accommodated by a decrease in inventories of 12.5. This unintended fall in inventories is an incentive to increase production to take advantage of higher than expected sales. The incentive exists until producers increase output and income to 250. Aggregate expenditure rises as income increases until equilibrium is reached at $Y = 250$. At that point actual inventory investment meets producer plans.

The change in autonomous expenditure by 25 caused an increase in equilibrium income by 50. The multiplier, defined as the change in equilibrium income caused by a change in autonomous expenditure:

$$\Delta Y / \Delta A = 2$$

You can construct a diagram like Figure 18.8 using the numerical values for GDP and aggregate expenditure in Table 18.4 and a 45° line to show equilibrium $Y_e = 200$. Example Box 18.2 at the end of the chapter illustrates the multiplier effect of a change in autonomous expenditure on equilibrium income using simple algebra.

The size of the multiplier

The multiplier is a number that tells us how much equilibrium output changes as a result of a change in autonomous expenditure. The multiplier is bigger than 1 because a change in autonomous expenditure changes income and sets off further changes in induced expenditure. The marginal propensity to spend on domestic output $(c-m)$ determines the induced expenditure.

In more general terms, because $mpc - mpm = c - m$ is the slope of the AE function, we can write:

$$\text{Multiplier} = \frac{1}{(1-\text{slope of AE})} \tag{18.4}$$

In the example in Table 18.4 with a induced expenditure $(c-m)Y = 0.5Y$ which also the slope of AE the multiplier is:

$$\frac{\Delta Y}{\Delta A} = \frac{1}{(1-\text{slope of AE})} = \frac{1}{(1-0.5)} = 2.0$$

But at a different time or in a different economy the marginal propensity to consume might be higher and the marginal propensity to import might be lower, for example let $c_1 = 0.85$ and $m_1 = 0.10$. Then induced expenditure on domestic output would be $(c_1 - m_1) = (0.85 - 0.10) = 0.75$ as would the slope of AE_1. Now a change in autonomous expenditure raises income and causes a higher change in induced expenditure. As a result the multiplier for this economy would be:

$$\frac{\Delta Y}{\Delta A} = \frac{1}{(1-\text{slope of AE})} = \frac{1}{(1-0.75)} = \frac{1}{0.25} = 4.0$$

As a result, national incomes in economies with higher rates of induced expenditure on domestic output will fluctuate more; experience stronger business cycles, as a result of changes in autonomous expenditure like investment and exports.

In the basic model we have here, the slope of AE is $(mpc - mpm)$, but this simple formula for the multiplier will still be useful when we introduce the tax system in the government sector in Chapter 19.

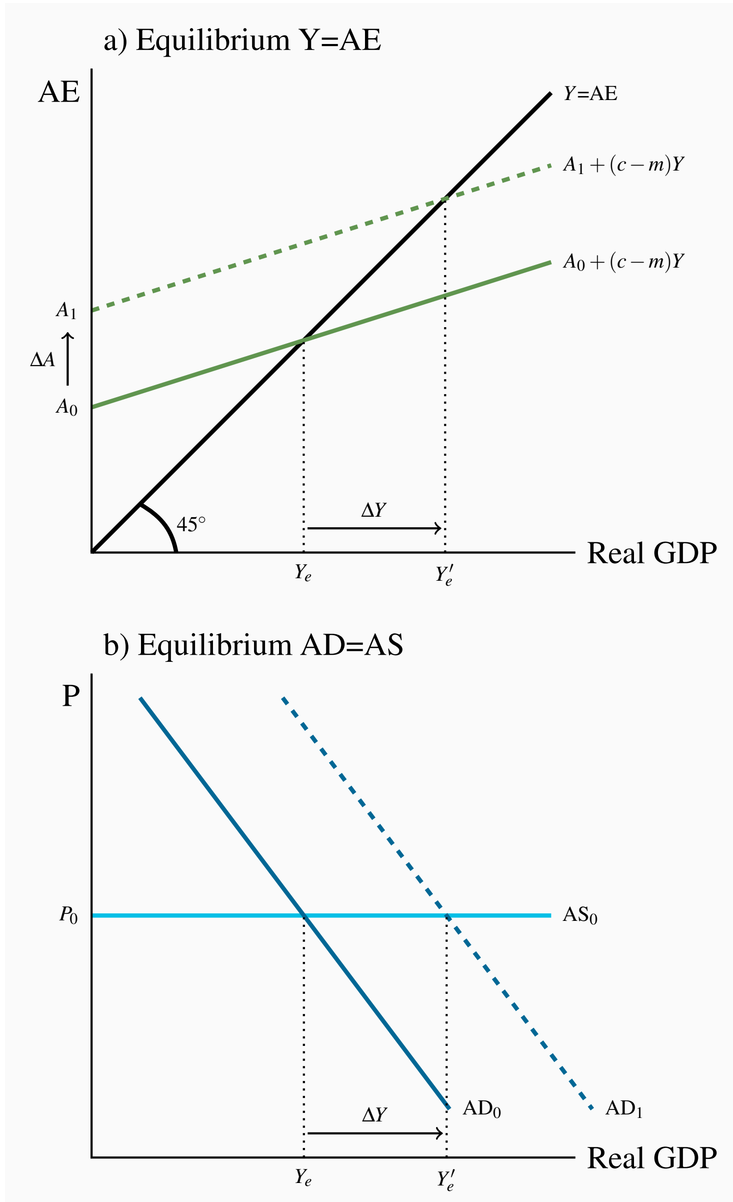
18.5 Equilibrium output and aggregate demand

In Chapter 17 and at the beginning of this chapter, we used an aggregate demand and aggregate supply model to explain business cycle fluctuations in real GDP and employment. In this chapter we have developed a basic explanation for the shifts in AD that cause changes in real output. In the short run:

- wages, prices, money supply, interest rates and exchange rates are assumed to be constant;
- distinction between autonomous and induced expenditures is important; and

- equilibrium real GDP requires output equal to planned aggregate expenditure.

Figure 18.9: Equilibrium GDP and aggregate demand



Equilibrium in a) determined the *position* of the AD curve in b). A *change in autonomous* expenditure in a) changes equilibrium Y and *shifts* AD by the change in autonomous expenditure times the multiplier.

In this model, investment and exports are the main sources of fluctuations in autonomous expenditures. The marginal propensities to consume and import describe the changes in aggregate expenditure caused by changes in income. These induced expenditures are the source of the multiplier. When business changes its investment plans in response to predictions and expectations about future markets and profits, or exports change in response to international trade conditions, the multiplier translates these changes in autonomous expenditure into shifts in the AD curve. Shifts in the AD curve cause changes in equilibrium output and employment.

Figure 18.9 shows how this works. Equilibrium real GDP in the upper panel determines the position of the AD curve in the lower panel.

Initially, equilibrium real GDP at the price level P_0 is determined by the equilibrium condition $Y_e = A_0 + (c - m)Y$ in the upper panel and by the equilibrium condition $\text{AD}_0 = \text{AS}_0$ in the lower panel.

Changes in autonomous expenditure shift the AD curve. If autonomous expenditure increased from A_0 to A_1 as shown in panel (a), equilibrium output would increase from Y_0 to Y_{e1}. The change in equilibrium output would be ($\Delta A \times$ multiplier). The AD curve would shift to the right to AD_1 as a result of the increase in autonomous expenditure. The size of the horizontal shift would be ($\Delta A \times$ multiplier).

This model provides an important first insight into the sources of business cycles in the economy. However, it is a pure private household/private business sector economy. Autonomous consumption, investment, exports and imports, and the multiplier drive real GDP and income and fluctuations in those measures of economic activity. There is no government, and thus no way for government policy to affect real output and employment. There is no financial sector to explain the interest rates and foreign exchange rates that affect expenditure decisions, and thus no monetary policy. In the next few chapters we extend our discussion of aggregate expenditure and aggregate demand to include the government sector and financial sectors, as well as fiscal and monetary policy. The framework becomes a bit more complicated and realistic, but the basic mechanics are still those we have developed in this chapter.

Nonetheless, this basic model explains why, in the last half of 2018 both Steven Poloz, the Governor of the Bank of Canada and Bill Morneau, the Minister of Finance, and many economic forecasters were concerned that low oil prices and exports, lower business investment expenditure, and unsettled international trading conditions might push the Canadian economy from slower growth and a recession in early 2019.

Figure 18.5 above shows that export growth did not give the hoped for increase in aggregate demand but investment expenditures did rise to offset some of that weakness in aggregate demand. Recognizing these economic conditions the Minister of Finance, Bill Morneau, tabled a mildly expansionary fiscal policy program in his 2018 budget. That policy planned for government budget deficits to finance expenditures on major infrastructure to increase the government component of aggregate demand and sustain growth in real GDP.

COVID-19 changed all this. Business shutdowns and limitations on personal activities brought larger cuts in autonomous expenditures, outputs than could be offset by new fiscal policy programs. In terms of AD, the curve shifted sharply to the left. That caused a large drop in real GDP.

Example Box 18.1: The algebra of the basic income – expenditure model.

The basic model has two components:

1. Aggregate expenditure = Autonomous expenditure plus induced expenditure:

$$\text{AE} = A_0 + (c - m)Y (18.2)$$

2. Equilibrium condition: GDP = Aggregate expenditure

$$Y = \text{AE}$$

Then equilibrium Y is found by substituting 1 into 2:

$$\begin{aligned} Y &= A_0 + (c-m)Y \\ Y - (c-m)Y &= A_0 \\ Y[1-(c-m)] &= A_0 \\ Y &= A_0/[1-(c-m)] \end{aligned}$$

Recall that $(c - m)$ is the slope of the AE function (i.e. $\Delta\text{AE}/\Delta Y$). Then in equilibrium:

$$Y = A/(1 - \text{slope of AE})$$

Using the numbers in the example in Table 18.3 $\text{AE} = 100 + 0.5Y$ and for equilibrium:

$$\begin{aligned} Y &= 100 + 0.5Y \\ Y - 0.5Y &= 100 \\ Y &= 100/(1-0.5) \\ Y &= 200 \end{aligned}$$

Example Box 18.2: The multiplier in a basic algebraic model.

Initial conditions in a basic model:

$$\begin{aligned}\text{Aggregate expenditure: } & \text{AE} = 100 + 0.5Y \\ \text{Equilibrium condition: } & Y = \text{AE}\end{aligned}$$

Then equilibrium national income is:

$$\begin{aligned}Y &= 100 + 0.5Y \\ 0.5Y &= 100 \\ Y &= 200\end{aligned}$$

Suppose autonomous expenditure increases by 25 to $A_1 = 125$:

$$\begin{aligned}\text{Aggregate expenditure: } & \text{AE}_1 = 125 + 0.5Y \\ \text{Equilibrium condition: } & Y = \text{AE}_1\end{aligned}$$

Then the new equilibrium national income is:

$$\begin{aligned}Y &= 125 + 0.5Y \\ 0.5Y &= 125 \\ Y_1 &= 250\end{aligned}$$

The change in autonomous expenditure by 25 increased equilibrium national income by 50.

The **Multiplier** is defined as the change in national income (ΔY) divided by the change in autonomous expenditure (ΔA) that caused it. In this example the multiplier is:

$$\Delta Y / \Delta A = 50/25 = 2.$$

Notice the multiplier is also equal to $1/(1 - \text{slope of AE})$. In this example the slope of AE is the marginal propensity to spend on domestic output $(c - m) = 0.5$. The multiplier is:

$$\Delta Y / \Delta A = 1/(1 - \text{slope of AE}) = 1/(1 - 0.5) = 1/0.5 = 2.$$

Key Concepts

Aggregate demand determines real output (Y) and national income in the short run when prices are constant.

Aggregate demand: aggregate expenditure (AE) at different price levels when all other conditions are constant.

GDP(Y): the national accounts measure of the sum of actual expenditure and income in the economy.

Aggregate expenditure (AE): planned expenditure by business and households.

Induced expenditure: planned expenditure that is determined by current income and changes when income changes.

Marginal propensity to consume ($mpc = c = \Delta C/\Delta Y$): the change in consumption expenditure caused by a change in income.

Marginal propensity to import ($mpm = m = \Delta IM/\Delta Y$): the change in imports caused by a change in income.

Induced expenditure $(c-m)Y$: planned consumption and imports expenditures that change when income changes.

Autonomous expenditure (A): planned expenditure that is not determined by current income.

Aggregate expenditure (AE): the sum of planned induced and autonomous expenditure in the economy.

Short-run equilibrium output: Aggregate expenditure current output are equal ($Y = \text{AE}$).

Unplanned changes in business inventories: indicators of disequilibrium between planned and actual expenditures – incentives for businesses to adjust levels of employment and output (Y).

Multiplier $(\Delta Y/\Delta A)$: the ratio of the change in equilibrium income Y to the change in autonomous expenditure A that caused it.

Exercises for Chapter 18

Exercise 18.1 Suppose that in an economy with no government the aggregate expenditure function is: $\text{AE} = 50 + 0.75Y$.

(a) Draw a diagram showing the aggregate expenditure function, and indicate the level of planned expenditure when income is 150.

(b) In this same diagram, show what would happen to aggregate expenditure if income increased to 200.

(c) What are the levels of autonomous expenditure and induced expenditure at income levels of 150 and 200.

(d) In this same diagram show what would happen if autonomous expenditure increased by 20.

Exercise 18.2 Suppose the media predicts a deep and persistent economic recession. Households expect their future income and employment prospects to fall. They cut back on expenditure, reducing autonomous expenditure from 50 to 30.

(a) Re-draw the aggregate expenditure functions you have drawn in your diagrams for Exercise 18.1 to show the effects, if any, of this change in household behaviour.

(b) Suppose the negative economic forecast also reduces induced expenditure in the economy from $0.75Y$ to $0.5Y$. In a diagram show the effect would this have on the aggregate expenditure functions you have drawn.

Exercise 18.3 Construct a table showing autonomous, induced and aggregate expenditure at different income levels (Y) for an economy with autonomous expenditure of 105 and induced expenditure of $0.5Y$.

(a) Using numbers from your table draw a diagram showing the aggregate expenditure function AE. What is the intercept of this function on the vertical axis?

(b) What is the slope of the AE function, and what does the slope measure?

(c) Write the equation for the aggregate expenditure function for this economy.

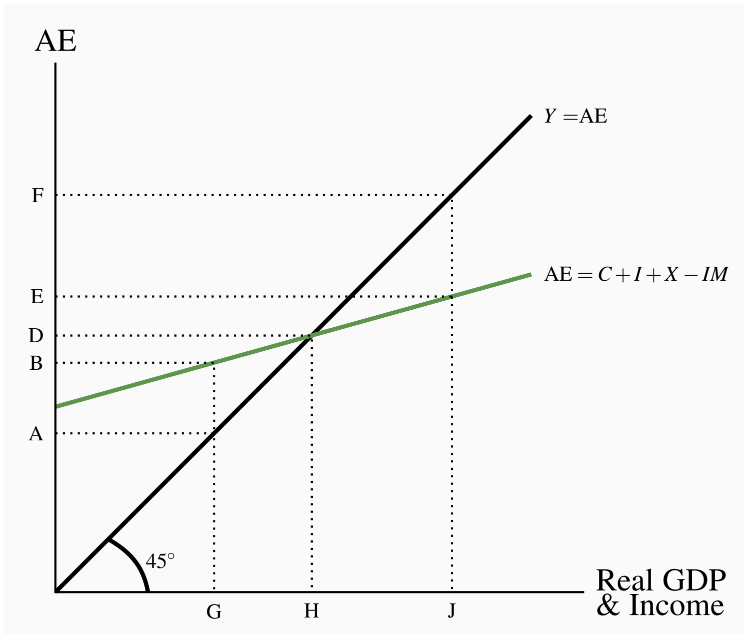
Exercise 18.4 Output and income are in equilibrium when planned expenditures AE are equal to national income, Y, in other words, meaning $Y = \text{AE}$.

(a) Suppose the AE function is $\text{AE} = 175 + 0.75Y$. Draw a diagram showing the aggregate expenditure function.

(b) In your diagram draw the 45° line that shows all points at which national income and aggregate expenditures are equal ($Y = \text{AE}$).

(c) Using your diagram, or a numerical example, or an algebraic solution, find equilibrium output and income in this example and show it in the diagram.

Exercise 18.5 The diagram below shows the aggregate expenditure schedule for the economy and the equilibrium condition on the 45° line.



(a) Suppose output is 0G. What is the level of planned aggregate expenditure? Is planned expenditure greater or less than output?

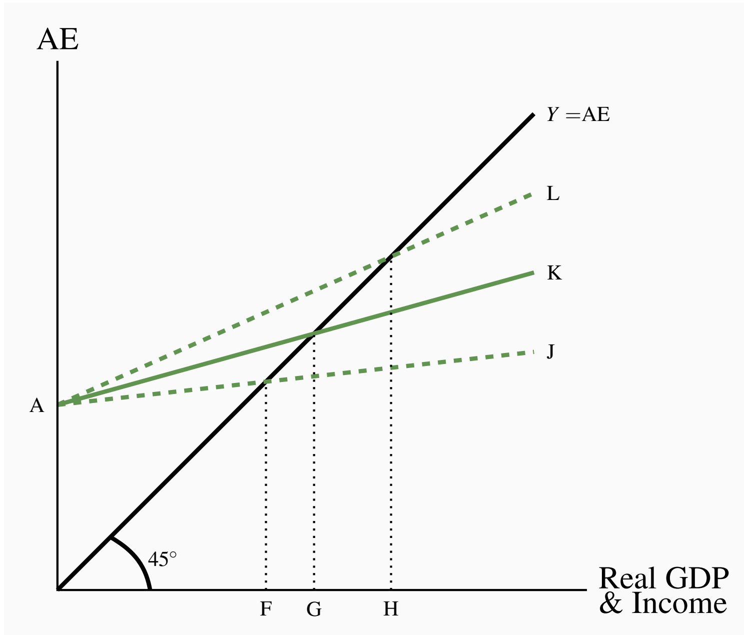
(b) What is the size of the unplanned change in inventories at output 0G?

(c) How will business firms respond to this situation?

(d) What is the equilibrium income and expenditure?

(e) Suppose output is at 0J: What is there an unplanned change in inventories?

Exercise 18.6 The following diagram shows an economy that initially has an aggregate expenditure function AK.



(a) What is the initial equilibrium real GDP?

(b) Suppose there is an increase in the marginal propensity to import. What is the new aggregate expenditure function?

(c) What is the new equilibrium real GDP and income?

(d) Suppose, instead, the marginal propensity to consume has increased. What is the new aggregate expenditure function? What is the new equilibrium real GDP and income?

Exercise 18.7 The distinction between autonomous and induced expenditure is important for the determination of equilibrium real GDP. Assume that the marginal propensity spend on domestic output is 0.70 and autonomous aggregate expenditure is zero.

(a) What is the equation for the aggregate expenditure function under these assumptions?

(b) Draw the aggregate expenditure function in an income-expenditure 45° line diagram.

(c) What is the equilibrium level of real GDP illustrated by your diagram?

(d) Explain why this is the equilibrium level of real GDP.

Exercise 18.8 Suppose the slope of the AE function is 0.6. Starting from equilibrium, suppose planned investment increases by 10.

(a) By how much and in what direction does equilibrium income change?

(b) How much of that change in equilibrium income is the result of the change in induced expenditure?

(c) How would your answers to (b) differ if the slope of the AE function was 0.8?

Exercise 18.9 Suppose autonomous expenditure is 100 and there is no induced expenditure in the economy.

(a) Write the aggregate expenditure function for this economy.

(b) Draw the aggregate expenditure function and the 45° line in a diagram.

(c) What is the equilibrium level of real output and income?

(d) By how much would equilibrium real output change if autonomous expenditure increased to 125? Show the change in expenditure and equilibrium in your diagram for part (b).

(e) What is the size of the multiplier? Explain your answer.

Chapter 19

The government sector

In this chapter we will explore:

Federal and provincial government budgets generate a lot of media coverage and public debate. In broad terms these debates involve the government's effectiveness and efficiency in providing public services and the government's fiscal management of the economy based on:

- the size of government and the composition of government expenditures,
- the tax system and policies used to finance expenditures,
- the budget deficit or surplus, and
- the size of the public debt, absolutely and relative to national income.

Consider recent examples.

In 2014 Minister of Finance Joe Oliver, in his 2014-15 Budget Plan, delivered on the Harper Government's earlier promise to balance the budget in 2015. But politics aside there really is no magic to a balanced budget. A balanced budget does not reduce the public debt. Nor does a balanced budget directly increase economic activity in an economy operating with a recessionary gap. Perhaps more interestingly, the government's actual budget balance at the end of its budget year depends importantly on the growth and level of GDP. As it turned out annual growth in real GDP fell to 0.6 percent and the unemployment rate increased to 7.1 percent. By March 31, 2015; the end of the fiscal year 2014-15, there was a small budget deficit despite increases in tax revenue and much smaller increase in expenditures. The public debt as measured by the accumulated deficit continued to increase.

Expenditure restraint to sustain balanced budgets and control public debt versus budget deficits to support economic growth and employment were major issues during the 2015 federal election campaign and debates.

During that campaign the Liberal Party proposed a sharp change in budget policy shifting from fiscal restraint to expansion. Citing the low growth rates in the economy and an increase in the unemployment rate they promised to stimulate the economy by running budget deficits for three years to finance major infrastructure projects. After the Liberal party won the election, Minister of Finance Bill Morneau's budget plan of April 2016 forecast the first of these deficits. The underlying argument was that increased government expenditure would lead to increased GDP and employment. Fiscal policy was then focused on improvements in economic activity in contrast to the previous governments focus on balancing the budget. Annual Budget deficits increased from \$2.9 billion in 2015-16; to \$19 billion 2017-18. Annual growth in real GDP increased from 0.8 percent in 2016 to 3.5 in 2017 and 2.2 percent in 2018. Unemployment rates fell from 7.1 percent to 5.6 percent. But the promise to limit annual deficits to \$10 billion for three years did not hold. The emphasis shifted to controlling the debt to GDP ratio. Fiscal policy, budget deficits vs surpluses and debt ratios, was one major economic issue in the 2019 federal election campaign.

This chapter examines these issues and the role of government in the economy. A government sector adds important new linkages and feedback effects to the basic model of Chapter 18. Government expenditures on goods and services (G), determined by government policy, are added to other autonomous expenditures. They affect AE and equilibrium income through the multiplier. Government revenue collected by a net tax rate (t) on national income reduces the income households have to spend on consumption and reduces induced consumption expenditure. The multiplier is reduced, reducing equilibrium national income and the effects of changes in autonomous expenditure on equilibrium national income.

The linkages and feedbacks work in both directions. Changes in government expenditures and taxes have effects on equilibrium national income. Government can use these effects to manage aggregate expenditure and equilibrium income. However, changes in equilibrium national income caused by changes in expenditure in other parts of the economy change incomes, government revenues and budget balances. As a result, changing economic conditions often lead to government budget outcomes that differ from initial targets and projections as in recent years.

Furthermore, over time governments must manage their budgets in ways that control the size of their debt relative to GDP. The fiscal policy governments implement through their budgets has dual objectives: Manage aggregate demand and, over time, manage the size of the public debt relative to national income. Until recently, Canadian governments have been more concerned about government budget surpluses, deficits, and debt than about demand management when designing fiscal policy.

In early 2020 the COVID-19 pandemic and containment policies led to dramatic changes in government budgets and fiscal policy in Canada. Households and businesses needed extensive financial support in order to deal with the effects of business closings, sudden large increases in unemployment and ongoing costs in both sectors. Plans to control budget balances and debt ratios were displaced by urgent immediate needs for financial support. Preliminary estimates of the costs of this support for the federal government's budget are summarized later in this chapter.

To explain the role of government in macroeconomic analysis and policy, we start with a brief look at the data on the size of the government sector in Canada.

19.1 Government in Canada

The total government sector in Canada includes the federal, provincial, and municipal governments, as well as hospitals. Table 19.1 shows total outlays by the government sector in 2017. These totaled $851 billion. Of this total, 30.6 percent was expenditure on government employees and 21.4 percent for the goods and services that provided government services to Canadians. The remaining 48 percent was transfer payments to persons, business, and non-residents, and interest paid on the outstanding public debt.

Table 19.1: Total government expense in Canada, 2017

Total Expense (billions $)	Compensation of employees %	Use of Goods & Services %	Consumption of Fixed Capital %	Subsidies & Grants %	Social Benefits %	Other Expense %	Interest %
851.1	30.6	21.4	8.2	2.3	0.6	24.9	7.3

Source: Department of Finance, *Fiscal Reference Tables, 2018*, Table 34

To provide some perspective, Table 19.2 compares the size of the government sector in Canada, relative to GDP, with the average for the G7 group of industrial countries (Canada, the United States, Japan, the United Kingdom, Germany, France, and Italy) in 2007 and 2017. These data illustrate two aspects of recent government budget activity that are of particular interest. The first is the size of the government sector in each country as measured by revenue, expenditure, budget balance and net public debt, all reported as a percent of GDP. The second is the change in government sector finances from 2007 to 2017, the period of the financial crisis, recession and prolonged recovery.

On the first point the 2007 data show expenditures by Canada's government sector—the combined federal, provincial, and local governments—on goods, services, and transfers were 38.6 percent of GDP. This was slightly more than the average for G7 countries. The difference reflects national political choices about the role the government sector plays in the economy.

Table 19.2: The general government sector in Canada vs. the G7 countries

	Total Revenues % GDP		Total Outlays % GDP		Budget Balance % GDP		Net Public Debt % GDP	
	2007	**2017**	**2007**	**2017**	**2007**	**2017**	**2007**	**2017**
Canada	40.4	39.3	38.6	40.3	1.8	−1.0	22.1	27.8
G7 Average (1)	36.0	36.0	38.3	39.3	−2.2	−3.4	55.7	87.5

Source: IMF. *World Economic Outlook* database and *Fiscal Monitor* (April 2018). *as reported in:*
Canada: Department of Finance, *Fiscal Reference Tables, 2018*, Tables 51-54.
(1) G7 average weighted by nominal GDP converted to U.S. dollars at average market exchange rates.

In 2007, Canada differed from the G7 average in terms of their government sector budget balances. Canada operated with a budget surplus (revenues were greater than expenditures), while the other countries had budget deficits. Canada's budget surplus was the latest in a series of annual government-sector budget surpluses over the period from 1997 to 2007. These budget surpluses reduced the outstanding public debt and reduced Canada's ratio of net public debt to GDP well below the average to the lowest in the G7.

The shift in fiscal conditions in the G7 from 2007 to 2017 was dramatic. The recession that followed the financial crisis of 2008 reduced employment and incomes in all countries. A period of persistent budget adjustments followed. Government revenue in Canada and the G7 fell relative to GDP. At the same time governments maintained or increased expenditures to stimulate demand, and some provided financial bailouts to banks to limit the impact of the financial crisis on bank balance sheets. In combination, these fiscal policy actions increased average G6 government outlays. Fiscal austerity followed as most countries tried to manage increases in their public debts. Even 10 years later, in 2017, budget deficits were up 1.2 percent of GDP. These budget deficits combined with slow growth in GDP, raised the average net public debt ratio more than 30 percentage points. Initially Canada also experienced a rise in the net public debt ratio but improved GDP growth and expenditure restraint reduced the ratio slightly by 2015. The government debt crisis that followed in several countries has dominated European economic conditions and policy debates and remained unsolved in 2019 in several countries.

19.2 Government expenditure & taxes

A basic government budget has two components:

1. A plan for *government expenditures* on goods and services, G.
2. A *net tax rate* on income, t, set to generate revenue to finance expenditure.

When added to the definition of autonomous expenditure, A is expanded to include **government expenditure** G.

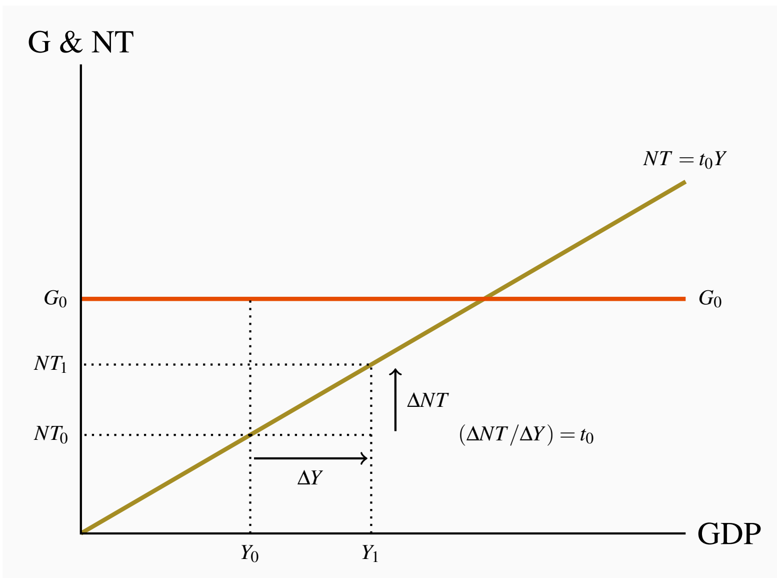
Government expenditure (G): government spending on currently produced goods and services.

We illustrate *autonomous* government expenditure in the same way we did with other autonomous expenditures, using a simple equation. For a specific level of government expenditure:

$$G = G_0 \tag{19.1}$$

In Figure 19.1 with income on the horizontal axis and government expenditure and net tax revenue on the vertical axis, a horizontal line intersecting the vertical axis at G_0 illustrates a particular level of government expenditure. Any change in government expenditure would shift this line up or down in a parallel way.

Figure 19.1: Government expenditure and net tax revenue function



The government raises revenue by levying direct taxes on incomes and pays out transfer payments such as old age security, employment insurance benefits, social assistance and interest on the public debt. The difference between taxes collected and transfers paid is **net taxes (NT)**, the net revenue collected by government from households.

Net taxes: taxes on incomes minus transfer payments.

Net taxes revenues relative to national income are also illustrated in Figure 19.1. With no autonomous net taxes and a tax rate t_0, the tax revenue line starts at the origin, $NT = 0$ when $Y = 0$, and rises as income rises, moving to the right in the diagram. The slope of the NT line is the net tax rate:

$$\Delta NT/\Delta Y = t$$

Figure 19.1 illustrates another important point. It shows that for a given level of planned government expenditure and a given tax rate, government expenditure will be greater than tax revenue at lower income levels and lower than tax revenue at higher income levels. There is a strong direct linkage between government finances and national income.

There is another but slightly more complex link between government taxes and non-government expenditures. The net tax rate on income reduces *induced* consumption. Net taxes reduce **disposable income**—the amount available to households for spending or saving—relative to national income. If YD **is disposable income**, Y national income and output, and NT net taxes:

$$\text{Disposable income} = \text{national income minus net tax revenue}$$

$$YD = Y - NT \tag{19.2}$$

Disposable income (YD): national income minus net taxes.

Suppose taxes net of transfers are about 15 percent of national income. The net tax rate $t = 0.15$. If national income Y increases by \$1, net tax revenue will increase by \$0.15, but household disposable income will increase by only \$0.85.

For simplicity, suppose the marginal propensity to consume ($mpc = c$) out of disposable income is 0.8. The consumption function is:

$$\text{Consumption} = \text{autonomous consumption} + \text{induced consumption based on disposable income.}$$

Then if:

$$C = 20 + 0.8YD$$

And the net tax rate is $t = 0.15$, disposable income is $(1 - 0.15) = 0.85$ times national income. Thus, consumption expenditure based on national income is:

$$C = 20 + 0.8 \times 0.85Y$$
$$C = 20 + 0.68Y$$

A change in national income of \$1 changes consumption expenditure by only 0.8 times, $(1 - t)$ a dollar. If the net tax rate is 0.15, consumption expenditure changes by only $\$1 \times (0.8 \times 0.85) = \$0.68$. Each extra dollar of national income increases disposable income by \$0.85, out of which households plan to spend 68 cents and save 17 cents. Table 19.3 gives a numerical example.

Table 19.3: A numerical example

a)	$C = 20 + 0.8YD$				b)	$C = 20 + 0.8YD$			
	$NT = 0$					$NT = 0.15Y$			
	$YD = Y$					$YD = (1 - 0.15)Y$			
						$C = 20 + 0.8(1 - 0.15)Y$			
a)	Y	NT	YD	C_1	b)	Y	NT	YD	C_2
	100	0	100	100		100	15	85	88
	300	0	300	260		300	45	255	224
	500	0	500	420		500	75	425	360

In the absence of taxation, in part (a) of the table, national income Y and disposable income YD are the same. The consumption function C_1 shows how much households wish to consume at each level of national income, based on the numerical example.

With a proportional net tax rate of 0.15, households still consume \$0.80 of each dollar of disposable income. Now part (b) shows YD is now only 0.85 of Y when the net tax rate is 0.15. Households consume only $0.8 \times 0.85 = \$0.68$ of each extra dollar of national income.

The effects of both government expenditure and taxation

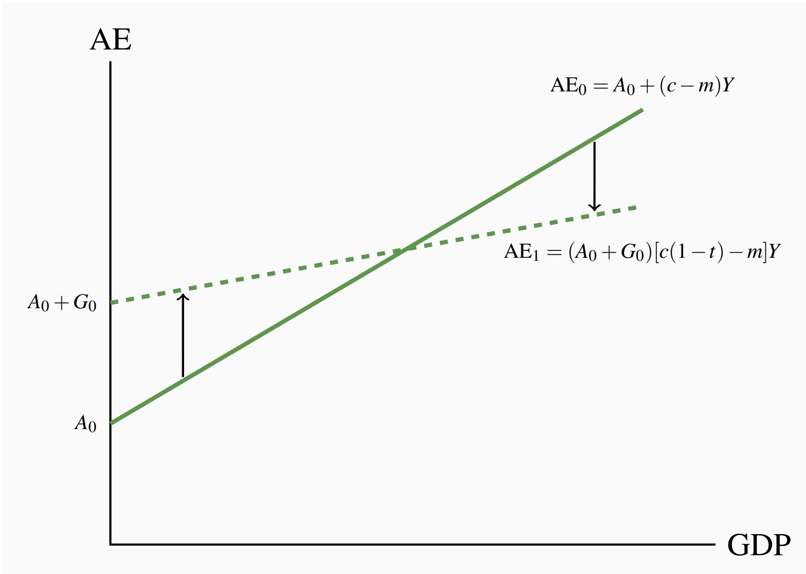
Aggregate expenditure includes both autonomous and induced expenditure. The net tax rate diverts some of national income to the government budget and lowers the disposable income on which induced expenditure is based. Net Taxes and imports both lower the marginal propensity to spend on domestic output. With the marginal propensity to consume out of disposable income (*mpc*), the marginal propensity to import (*mpm*) and the net tax rate (t), the slope of the AE function, the marginal propensity to spend on domestic output is:

Slope of AE $=$ marginal propensity to consume disposable income $-$ marginal propensity to import

$$\frac{\Delta \text{AE}}{\Delta Y} = c(1 - t) - m \tag{19.3}$$

Figure 19.2 shows the combined effect of government expenditure and net taxes on the aggregate expenditure function.

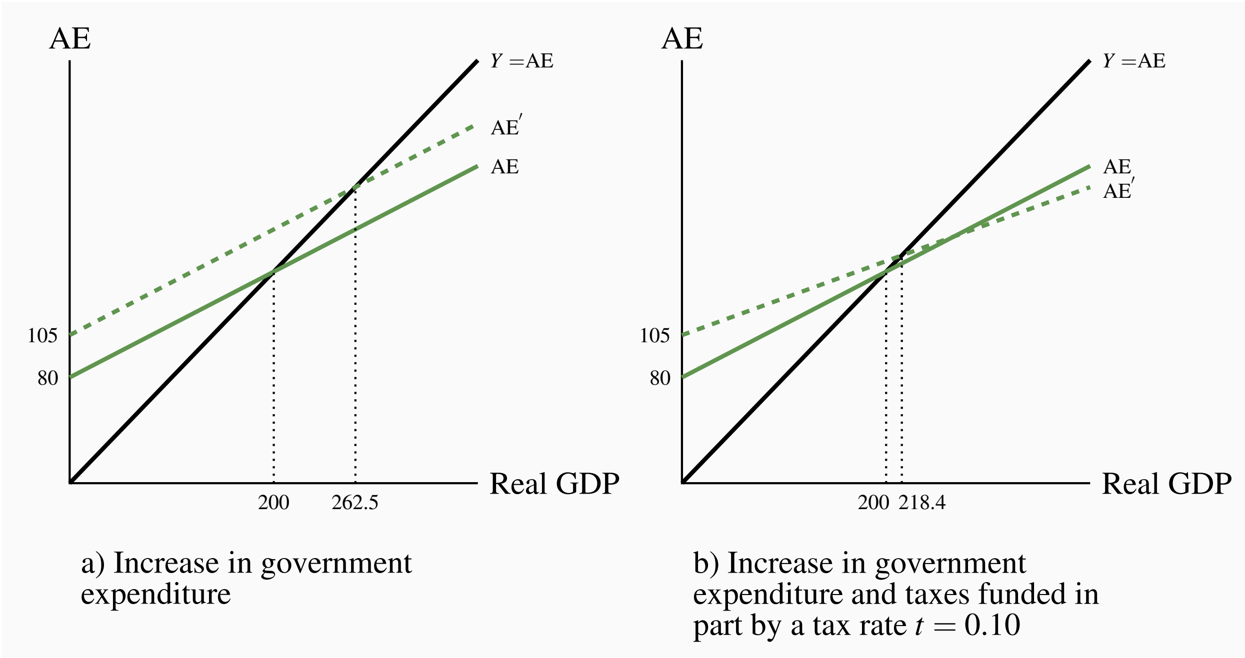
Figure 19.2: The effect of G and NT on AE



New government expenditure and taxes shift the AE function up by G_0 and lower its slope from $(c-m)$ to $c(1-t)-m$. A change in G shifts AE vertically. A change in t changes the slope of AE.

As in the basic model, equilibrium real GDP equals autonomous aggregate expenditure multiplied by the multiplier. The government sector adds a new autonomous expenditure component G_0 that shifts the AE function, and a new induced expenditure factor, the net tax rate (t), which lowers the slope of the AE function. Figure 19.3 illustrates the effects on equilibrium GDP and the multiplier.

Figure 19.3: Government expenditure, taxes, and equilibrium real GDP



Part a) of Figure 19.3 shows the effect of adding government expenditure to aggregate expenditure in the basic model. Government expenditure $G = 25$ is autonomous and shifts the AE curve up to a new intercept at 105. With $mpc = 0.8$ and $mpm = 0.2$ the slope of AE is 0.6 and the multiplier is 2.5. Equilibrium Y increases by $25 \times 2.5 = 62.5$ to the new equilibrium 262.5. An increase in autonomous government expenditure increased equilibrium GDP. A cut in government expenditure would reduce it. In this example, since the government has not introduced a tax system the government expenditure is financed by borrowing money through the sale of bonds.

Part b) shows the effect of government expenditure finance by net taxes on income. Government expenditure is 25 and tax revenues are raised by a net tax rate of 10 percent, $t = 0.10$ and $NT = 0.10Y$. Higher autonomous government expenditure raises the intercept of the AE function by 25 to 105. The net tax rate lowers the slope of the AE function and reduces the size of the multiplier from 2.5 to 2.08. Equilibrium GDP does increase to 218.4 because the government expenditure of 25 is greater than the net tax revenue (0.10×218.4) of 21.84. The government's budget is in deficit.

Example Box 19.1 at the end of the chapter gives the basic algebra of Figure 19.3.

The multiplier revisited

The multiplier relates changes in equilibrium national income to the changes in autonomous expenditures that cause them. The formula in Chapter 18 still applies.

$$\text{The multiplier} = \frac{1}{1 - \text{slope of AE}}$$

Without government and taxes, disposable income and national income are the same. With induced expenditure based on the marginal propensity to consume ($mpc = c$) and the marginal propensity to import ($mpm = m$) the multiplier is:

$$\frac{\Delta Y}{\Delta A} = \frac{1}{1 - c + m}$$

With government net taxes proportional to income, $NT = tY$, disposable income is less than national income namely $Y - tY$. This reduces the marginal propensity to consume out of national to $c(1-t)$, reducing the slope of the AE function.

$$\frac{\Delta \text{AE}}{\Delta Y} = c(1-t) - m$$

As a result, the multiplier, which is $1/(1 - \text{slope AE})$ is smaller:

$$\frac{\Delta Y}{\Delta A} = \frac{1}{1 - c(1-t) + m} \tag{19.4}$$

The numerical values in Figure 19.3 provide one example of this change in the multiplier.

Now that we have seen that government expenditure and net tax taxes have effects on aggregate expenditure and equilibrium income, it is time to examine the effects of government budgets on AE, AD, and real GDP. The government implements fiscal policy through its budget.

Application Box 19.1: The multiplier in Canada

The multiplier plays a key role in the AE and AD/AS model of the economy. But what is the size of the multiplier in Canada? A simple statistical estimate, using Statistics Canada annual data for real GDP and consumption expenditures, gives a Canadian marginal propensity to consume out of national income $c(1-t) = 0.54$, and marginal propensity to import $m = 0.34$. Using these estimates, we get a multiplier for Canada:

$$\Delta Y/\Delta A = 1/(1-0.54+0.34) = 1/(1-0.2) = 1.25$$

Recent data gives an estimate of the Canadian marginal propensity to consume out of disposable income of $mpc = 0.88$. If there were no taxes or imports, an $mpc = 0.88$ would mean a multiplier of about 8.33. The difference between the multipliers 1.25 and 8.33 shows clearly the automatic stabilization coming from the net tax rate and marginal propensity to import.

19.3 The government's budget function

It is important to make a clear distinction between a government budget as it affects and is affected by national income, and an individual or household budget. The large size of the government sector relative to the economy (see Table 19.2) means any change in government expenditure or taxes also changes national income and government tax revenue. An increase in government expenditure also increases its tax revenue by increasing national income. A cut in government expenditure reduces national income and tax revenue. A household can improve its budget balance by cutting its expenditure without affecting its income. A government cannot. Recent experience in many countries provides a record of the difficulties governments have in trying to reduce budget deficits by cutting government expenditures.

The government budget function provides a useful tool for the discussion of these questions. It illustrates two-way linkages between the government budget and national income and the internal feedback within the government budget.

Think of a budget as the revenue and spending plan of an individual, a company, or a government. The government budget describes what goods the government will buy, the public services it will provide during the coming year, what transfer payments it will make, and how it will pay for them. Most spending is financed by taxes, but some revenue comes from charges for services. A balanced budget has revenues equal to spending. When revenues exceed spending, there is a budget surplus and some of the outstanding public bond debt expires. When revenues fall short of spending, there is a budget deficit, which is financed by borrowing through the sale of government bonds to the public.

> **Government budget**: planned government spending and revenue.
>
> **Balanced budget**: revenues are equal to expenditures.
>
> **Budget surplus**: revenues are greater than expenditures.
>
> **Budget deficit**: revenues are less than expenditures.

Continuing to use G for government expenditure on goods and services, and NT for net tax revenue or taxes minus transfer payments (ignoring other sources of revenue, for simplicity):

$$\text{Government budget balance } (BB) = \text{Net tax revenue } (tY) - \text{government expenditure } (G)$$

$$BB = tY - G \tag{19.5}$$

The budget balance, whether deficit, surplus or zero, is determined by three things:

1. the *net tax rate* t set by the government;
2. the level of *expenditure* G set by the government; and
3. the level of *output* Y determined by AE and AD.

We can summarize the relationship between national income and the government's budget balance in a way we will find useful for discussing a government's budget and fiscal policies. Figure 19.4 shows a **budget function**. This budget function assumes $t_0 = 0.20$ and $G_0 = 200$. This is a specific fiscal program and budget plan, namely BB_0 where:

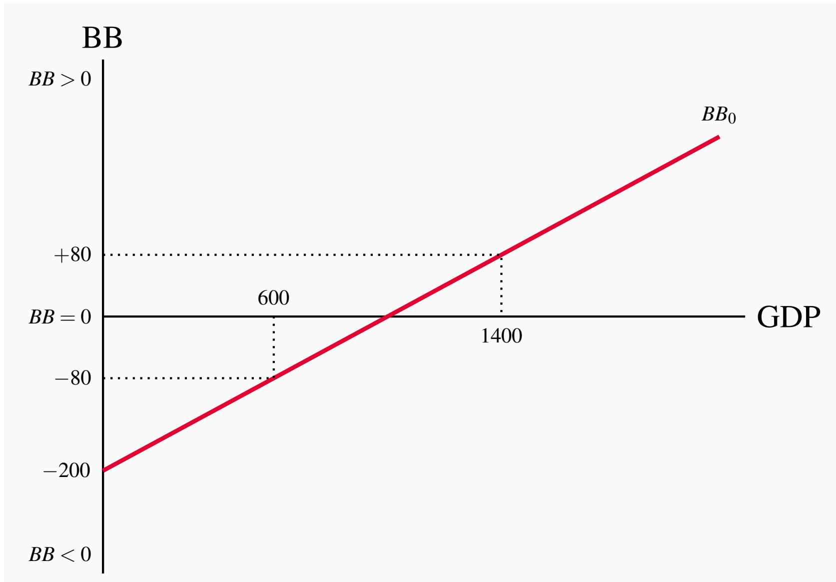
$$BB_0 = t_0 Y - G_0$$

that gives:

$$BB_0 = 0.2Y - 200$$

> **Budget function**: the relationship between the budget balance and the level of national income for a specific budget program.

Figure 19.4: The government budget function



The budget function is positioned by $G = 200$ with a slope $\Delta BB/\Delta Y = t = 0.2$. The budget balance depends on the equilibrium level of GDP determined by AE.

A budget function is a simple illustration of the way the budget balance for one fiscal policy program depends on levels of national income.

In Figure 19.4, if national income Y were 600 the budget BB_0 would be in deficit by 80. If a rise in autonomous aggregate expenditure increased equilibrium national income to 1400 the budget BB_0 would be in surplus. Changes in national income induce changes in the revenues generated by the tax rate $t = 0.2$ and automatically change the government's budget balance. The fiscal program and budget plan have not changed.

Once that fiscal program is set, the budget function is set, but the budget balance is not. The budget balance depends on the performance of the economy in terms of national income. In presenting the budget, the Minister of Finance gives a forecast of the budget balance based on a forecast of national income. If the income forecast is wrong, the budget program will result in either a larger or smaller budget balance than initially predicted.

As noted earlier, recent experiences illustrate this relationship. For example, in April of 2014 the Minister of Finance, Joe Oliver, tabled a government budget plan that was intended to result in a balanced budget ($BB = 0$) in the fiscal year 2014–15. That budget plan was based on the assumption the GDP in Canada would grow at an annual rate of 2 per cent in 2014–2015 and generate corresponding government tax revenue. However, by mid-July 2014 there were serious questions about that growth rate assumption. The Parliamentary Budget Office in a Budget Update that used the Bank of Canada's latest forecasts for real GDP growth at an annual rate of 1.0 percent estimated a budget deficit of \$1.5 billion for 2014–15. The promised balanced budget was at risk as real GDP would be less by the end of the budget year. The finance minister makes a budget plan but the performance of the economy determines the budget outcome.

Final results in the *2018 Budget* provide another example. That budget predicted a fiscal year deficit, ending in March 2019, of \$18.1 billion but the *2019 Budget* that followed reported a larger final deficit of \$19.0 billion. Part of the explanation for this larger deficit comes from the decline in income growth in the last part of 2018. A fall in annual GDP growth by 1% results in a \$4.8 billion fall in the annual budget balance based on the sensitivity analysis provided in the 2018 budget. www.budget.gc.ca/2018/docs/plan/toc-tdm-en.html

Policies used in early 2020 to contain the spread of the COVID-19 pandemic caused financial distress for many businesses, households and individuals. To provide financial support to offset this distress, the federal government introduced a long list of support measures to partially offset the loss of income and employment. The preliminary estimates of the effects on the government's budget balance and debt ratio are presented later in this chapter.

Finally, notice that because a budget function describes one fiscal plan, any change in the fiscal plan will change the *BB* line to show a new budget function. The differences between budget plans and outcomes described above were movements along budget functions set out in the annual budget documents.

19.4 Fiscal policy & government budget balances

Fiscal policy is the government's use of its taxing and spending powers to affect aggregate expenditure and equilibrium real GDP. The main objective of fiscal policy is to stabilize output by managing aggregate demand, keeping output close to potential output, and reducing the size and duration of business cycle fluctuations. This requires changes in the government's expenditure plans and tax policy to offset *predicted* changes in autonomous expenditures that would otherwise push the economy away from equilibrium at potential output.

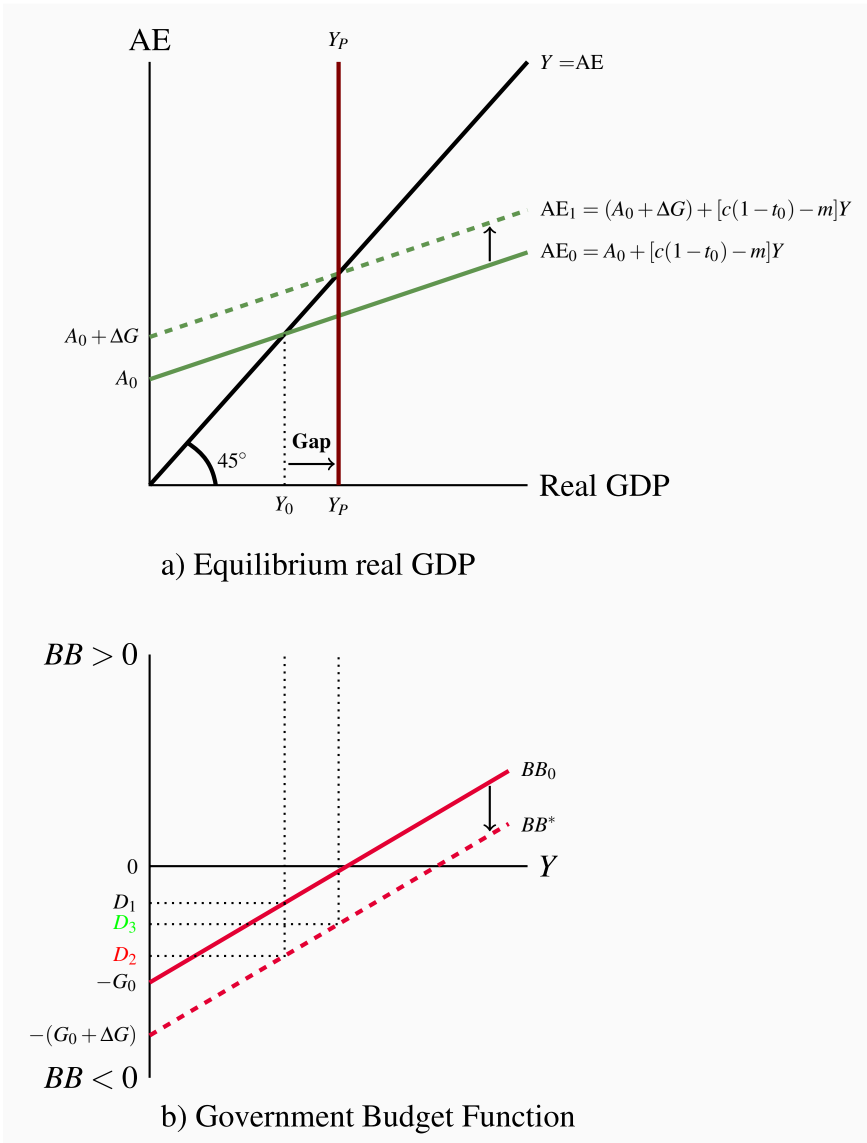
> **Fiscal policy**: government use of taxes and spending to affect equilibrium GDP.

In 2008 and 2009, for example, the international financial crises and the recession that followed led to fiscal stimulus programs in most industrial countries, like Canada's federal '*Economic Action Plan*', and calls for international coordination of fiscal stimulus. This fiscal stimulus led in turn to increased budget deficits and national debts and, especially in Europe, to national or sovereign debt crises.

Fiscal stimulus

Figure 19.5 illustrates the use of fiscal policy to eliminate an output gap. In the Part a) of the diagram the economy has a recessionary gap at equilibrium $Y_0 < Y_P$ because aggregate expenditure, AE, is not high enough to give equilibrium at Y_P. Government can intervene to raise AE to AE_1 by increasing government expenditures or by lowering the net tax rate or a combination of the two. In this example, the government chooses to increase G from G_0 to $G_0 + \Delta G$. The increase in G, working through the multiplier raises equilibrium real GDP to potential output.

Figure 19.5: Fiscal policies to eliminate an output gap



An increase in G reduces the budget balance and BB drops to BB^*, increasing the deficit from D_1 to D_2. The higher autonomous G increases AE to AE_1 and real GDP increases to Y_P, eliminating the output gap. That increase in Y increases tax revenue and offsets some of the initial increase in the deficit such that the final budget deficit is D_3.

Part b) of the diagram shows the change in the fiscal program used to provide fiscal stimulus and expand AE and equilibrium real GDP. Increased G lowers the budget function at every income level by ΔG to the new function BB^*.

Before the increase in G the budget had a small deficit D_1. The deficit rises initially to D_2 as a result of higher government expenditure and as yet no increase in income. Then, as the increased government expenditure raises AE and real GDP, the economy moves to the right along the budget function BB^* as higher national income generates higher tax revenue. Higher revenue offsets some of the increase in the deficit caused by ΔG leaving deficit D_3, larger than the deficit D_1 but less than the increase in spending and deficit used to stimulate the economy.

Fiscal policy makes changes in net tax rates and government spending that are intended to change aggregate expenditure and aggregate demand and stabilize equilibrium output at potential output. These changes change the government's budget function and the budget balance. In the case of fiscal expansion in Figure 19.5 the increase in equilibrium income means the fiscal stimulus is partly self-financed by the increase in equilibrium income.

Fiscal austerity

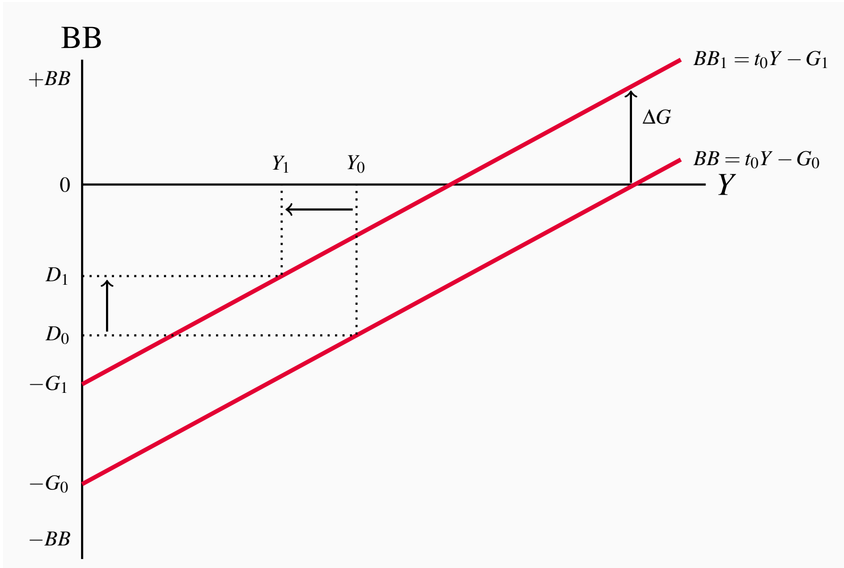
At times government directs fiscal policy not to the level of economic activity but to state of the government budget balance and public debt ratio. A balanced budget $(BB = 0)$ is often the target of a fiscal plan that involves some combination of cuts in government expenditure and tax increases to cut budget deficits. This is fiscal austerity. Federal government fiscal policy in Canada from 2010 to 2015 is one example. Also, several provincial governments have set reduced budget deficits and balanced budgets as targets to be met in their current mandates.

> **Fiscal austerity**: cuts in government expenditure and/or increases in taxes aimed at improving the government's budget balance.

But the same linkage and feedback effects that work to finance part of a fiscal expansion work to opposite effect in the case of fiscal austerity. If a government cuts expenditures in trying to reduce a budget deficit, some of the initial effect on the budget balance is lost to falling tax revenue. Lower G lowers equilibrium income and lower income lowers tax revenues. Larger budget expenditure cuts to make up for lost revenues make matters worse. This was the experience of many European countries in the years following the recession of 2008. Greece is the extreme case.

Figure 19.6 illustrates the effects of a cut in G designed to reduce a budget deficit.

Figure 19.6: The fiscal policy to reduce a budget deficit



A cut in government expenditure to reduce the budget deficit D_0 at Y_0 shifts the BB line up to BB_1. The deficit is reduced but the austerity policy lowers autonomous aggregate expenditure and equilibrium national income from Y_0 to Y_1. The net reduction in the deficit is less than the cut in G.

Because the observed budget balance combines autonomous (G) and induced (NT) components it is important to consider if the *observed budget balance*—whether surplus, balanced, or deficit—is a good measure of the government's policy action or fiscal stance.

Does the budget balance show whether fiscal policy is *expansionary*, aiming to raise national income, or *contractionary*, aimed at deficit control and reduction?

In itself, the budget balance may be a poor measure of the government's fiscal stance, because the budget balance can change for reasons unconnected to fiscal policy. Even if G and t are unaltered, a fall in investment or exports will reduce national income and output. In turn, this reduces net tax revenue and reduces the budget balance. Indeed, any change in non-government autonomous expenditure changes equilibrium income, net tax revenue, and the government's budget balance. Canadian experiences with volatile oil prices and lower energy export growth are good illustrations.

The *structural budget balance* provides a fiscal indicator that helps to solve this problem.

The structural budget balance

The **structural budget balance (*SBB*)** is an estimate of what the budget balance would be if the economy were operating at potential output. By evaluating the budget at a fixed level of income, namely potential GDP, the structural budget balance does not change as a result of business cycle fluctuations in output. In terms of the budget function we used above, the structural balance is:

$$SBB = tY_p - G \tag{19.6}$$

Structural budget balance (SBB): the government budget balance at potential output.

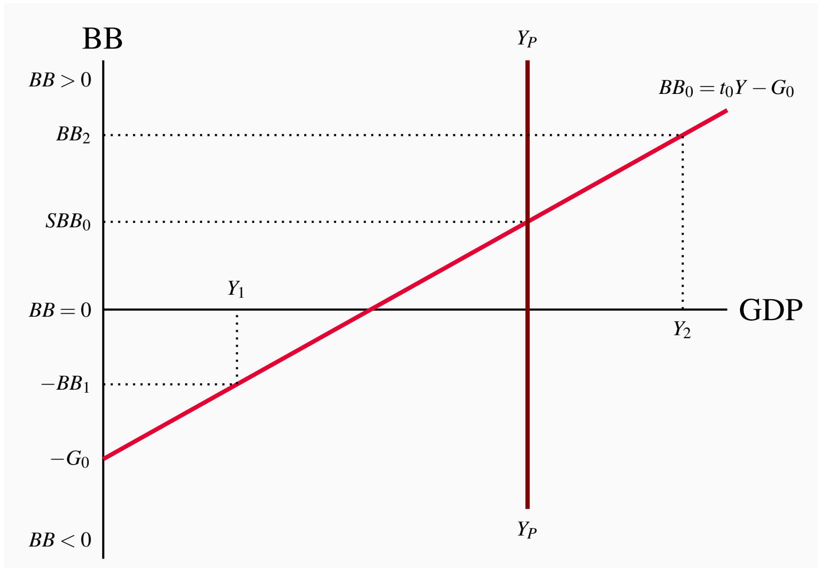
Notice that this structural budget function differs from the general budget function of Equation 19.5 by calculating net tax revenue at potential output, Y_P, rather than at any Y.

Using the previous numerical example, suppose government expenditure is 200 and the tax rate is 0.20. Holding these terms of the fiscal plan constant, as in Figure 19.4, the budget balance is a deficit at any income below 1000 and a surplus at any income above 1000. If, given other components of aggregate expenditure, the equilibrium output is 800, the actual budget balance will be a deficit. Net tax revenue will be $NT = 0.2 \times 800 = 160$. With government expenditure of $G = 200$, $BB = 160 - 200 = -40$.

Conversely, suppose higher AE makes equilibrium output 1200. With a tax rate of 0.20 and government expenditure of 200, the budget balance would be a surplus of 40. The important point of these examples is that we cannot tell the stance of fiscal policy, or a change in the stance of fiscal policy, by looking at the budget balance reported by the Department of Finance and published in the media. We need to look at a *structural budget balance*, calculated at potential output (Y_P) that is not changed by business fluctuations in actual output around potential output. Estimates of structural budget balances in Canada are published by the Department of Finance in *Fiscal Reference Tables*, `https://www.canada.ca/en/department-finance/services/publications/fiscal-reference-tables/2019.html`, Tables 17 and 46, as 'cyclically adjusted budget balances'.

Figure 19.7 illustrates the concept of the structural budget deficit as compared to the actual budget deficit. The fiscal policy program is $BB_0 = t_0Y - G_0$. If the economy were operating at potential output Y_P the budget balance would be SBB_0, the structural balance. At any other level of income, resulting for example from cyclical changes in autonomous investment or exports, the actual budget balance would be different from the structural budget balance. A fall in autonomous expenditure that lowered equilibrium GDP to Y_1 would result in a budget deficit – BB_1 in the diagram. There has been no change in fiscal policy – just a change in the actual budget caused by a business cycle change in national income.

Figure 19.7: Actual and structural budget balances



Structural budget balance $SBB_0 = t_0Y_P - G_0$. Actual budget balance $BB_1 = tY_1 - G_0$.

A change in the fiscal plan that changed the net tax rate or a change in planned government expenditure would change the budget function and the structural budget balance. There would be a different budget function line in the diagram. The slope of the line would increase with an increase in the tax rate or fall with a cut in the tax rate. A change in planned government expenditure would change the position of the line. In either case the new structural budget balance would indicate that fiscal policy had changed. The change in fiscal policy would increase or decrease aggregate expenditure.

19.5 Automatic and discretionary fiscal policy

Automatic fiscal policy

Automatic stabilizers have a great advantage. They are built into the budget program by setting the net tax rate, and work automatically. There is no need to determine if the shift in autonomous expenditure is transitory or persistent. By reducing the sensitivity of the economy to expenditure shocks, automatic stabilizers are always at work reducing the size of output and employment fluctuations.

> **Automatic stabilizers**: tax and transfer programs that reduce the size of the multiplier and the effects of *transitory* fluctuations in autonomous expenditures on equilibrium GDP.

Income taxes and transfers, such as unemployment benefits, are important automatic stabilizers. At given net tax rates, a fall in national income, output, and employment raises payments of unemployment benefits and reduces tax collections. Both effects mean that disposable income changes by less than the change in national income. The slope of the aggregate expenditure function $(c(1-t)-m)$ is lower, and so is the multiplier. Conversely, in a boom, net tax revenues rise and disposable income rises by less than the rise in national income, which helps dampen the boom.

However, automatic stabilizers only serve to moderate the fluctuations in real GDP caused by fluctuations in autonomous expenditure. They do not offset those autonomous expenditure disturbances. There is no automatic change in autonomous government expenditure or tax rates. Those changes usually come from discretionary fiscal policy.

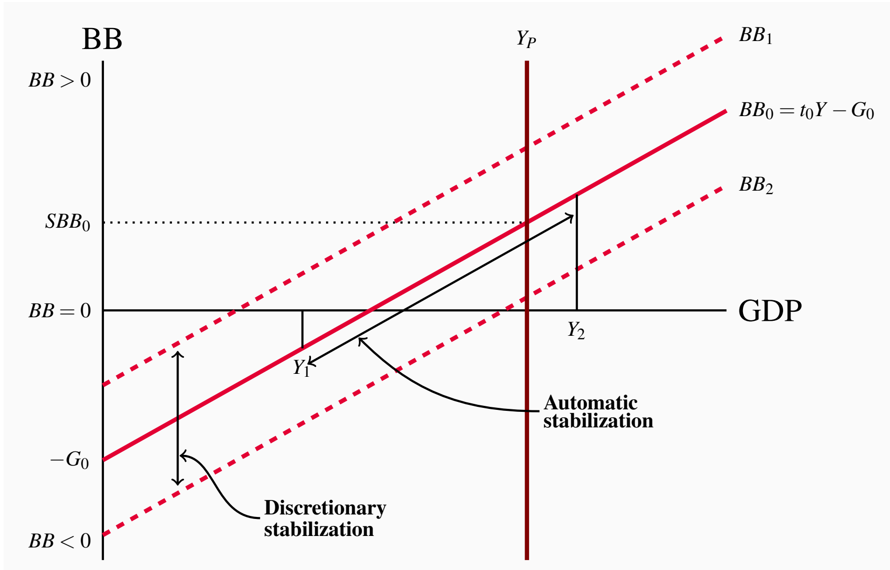
Discretionary fiscal policy

Governments use **discretionary fiscal policies** to offset *persistent* changes in autonomous expenditures. A persistent drop in investment or exports would be offset by an increase in government spending and by cutting taxes, or both as for example the Canadian government's *Economic Action Plan* in response to the recession of 2009. Another example is the planned annual deficit of $10 billion in the federal government's 2017 Budget, intended to finance major infrastructure investments and support increased aggregate demand, after the slowdown in growth in the previous years. Alternatively some provincial governments introduced austerity budgets in 2018 and 2019, cutting expenditures to try to reduce budget deficits.

> **Discretionary fiscal policy**: changes in net tax rates and government expenditure intended to offset *persistent* autonomous expenditure shocks and stabilize aggregate expenditure and output.

The budget function and the structural budget balance we discussed earlier provide a good illustration of automatic and discretionary fiscal policy. Figure 19.8 shows a government budget function $BB_0 = t_0Y - G_0$ and a structural budget balance SBB_0 at potential output Y_P. This budget function represents a fiscal program designed by the Minister of Finance and approved by parliament.

Figure 19.8: Automatic and discretionary fiscal policies



Automatic stabilization comes from changes in the budget balance along the BB_0 line as Y fluctuates between Y_1 and Y_2. *Discretionary stabilization* shifts the budget function as a result of changes in government expenditure or taxes.

Discretionary fiscal policy sets both the position and slope of the budget function. A change in discretionary policy would *change the entire budget line*. Figure 19.8 illustrates discretionary policy as shifting the BB line up to BB_1, in the case of restraint or austerity, or down to BB_2 to provide fiscal stimulus. Automatic stabilization is a part of all these programs. It comes from *the slope of the budget function*, the net tax rate t_0 in this case. Any fluctuations in private sector autonomous expenditures cause changes in income Y. These changes in Y for example, down to Y_1 or up to Y_2, cause *movements along the budget function* and a change in the budget balance, as shown in Figure 19.8. The effect of the change in the budget balance is stabilizing. A larger net tax rate would mean larger automatic changes in the budget balance in response to changes in income and more automatic stabilization.

When we use the budget function to show fiscal policy changes, we can also consider more complex programs that change both the slope of the function and the structural balance.

It is quite easy to present fiscal policy in theory and illustrate it in diagrams but does it work in the real world? If governments have fiscal tools to stabilize and offset fluctuations in aggregate expenditure and demand, why do we still experience business cycles, including the recession of 2009 and the prolonged recovery?

The answer has several dimensions. While automatic stabilizers moderate the severity of fluctuations in autonomous expenditures they do not offset those fluctuations. That calls for discretionary fiscal policy, namely a change in the budget plan involving changes in autonomous government expenditures and net tax rates. The process is partly economic and partly political and can take time.

The timelines involved are frequently defined in terms of recognition lags, decision lags, implementation lags and impact lags. It takes time to recognize a *persistent* shift in aggregate expenditure and identify its source. This involves the availability of economic data and economic analysis to establish the size and source of shift in economic conditions. That in turn provides the basis for the design of the new budget program required. The implementation of the new budget is a political process. It may involve substantial time and changes to the budget before it passes. Once the budget passes and new expenditure plans and tax rate are in effect it takes time for them to work through the economy and have their full impact on aggregate expenditure and national income. As a result, economic fluctuations are well underway before discretionary fiscal policies can shift to offset them. Discretionary policies may still provide stabilization but they do not completely eliminate business cycle fluctuations.

19.6 The public debt and the budget balance

Budget balances and outstanding debt are closely related. A student's debt at the time of graduation is the sum of her budget balances during years of study. In any year in which her income is less than her expenses, she finances the difference by borrowing. In another year, if income is greater than expenses, she can repay some previous borrowing. In the end, the sum of borrowings minus the sum of repayments is her outstanding student debt (loan). This debt is a result of borrowing to finance investment in education.

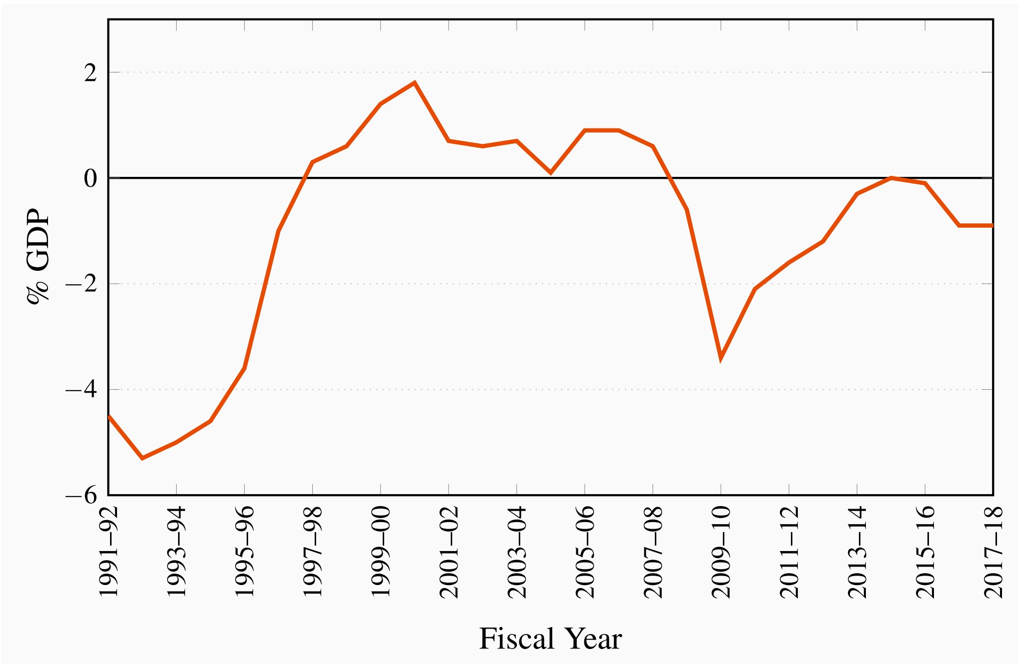
Similarly, the outstanding **public debt (*PD*)** at any point in time is simply the sum of past government budget balances. Governments borrow to finance budget deficits by selling government bonds to households and businesses. Budget surpluses reduce the government's financing requirements. Some bonds previously issued mature without being refinanced. In simple terms, the budget balance in any year changes the outstanding public debt by an equal amount but with the opposite sign. A positive balance, a surplus ($BB > 0$), reduces the public debt ($\Delta PD < 0$). A negative balance, a deficit ($BB < 0$), increases the public debt ($\Delta PD > 0$). Using PD to represent the outstanding public debt, we can express the link between the public debt and the government's budget balance as:

$$\Delta PD = -BB \tag{19.7}$$

> **Public debt (*PD*)**: the outstanding stock of government bonds issued to finance government budget deficits.

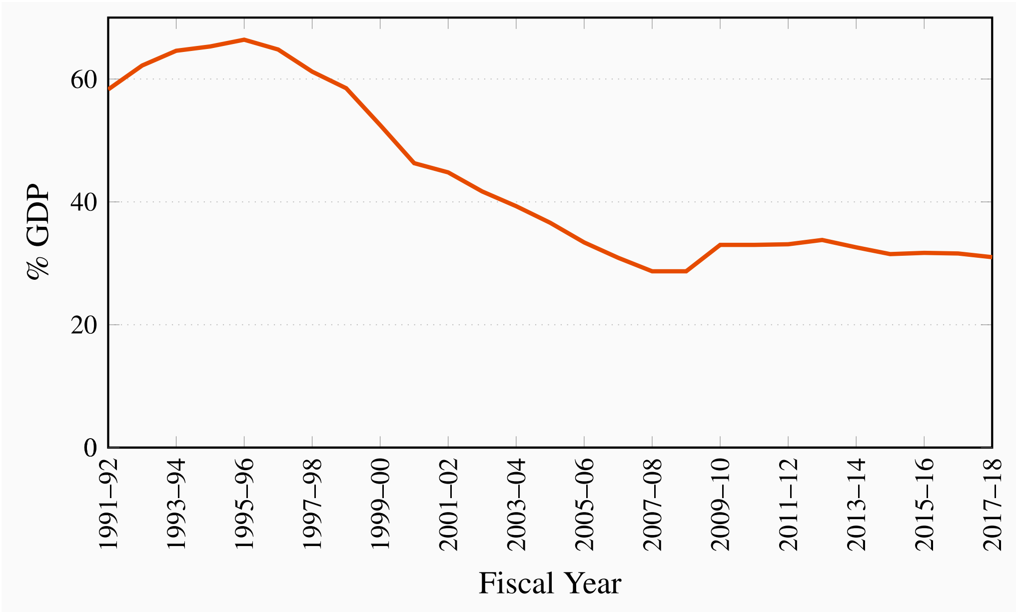
Figures 19.9 and 19.10 show the relationship between the government budget balance and the change in the public debt relative to GDP based on Canadian data for the 1991–92 to 2017–18 period. Recognizing that growth in the economy makes absolute numbers for deficits and debt hard to evaluate, the budget balance and the change public debt are presented as percentages of nominal GDP. The effects of budget balances on the public debt are illustrated clearly in the diagrams.

Figure 19.9: Government of Canada Actual Budget Balances
1991–92 to 2018–19
% Potential GDP



Source: Department of Finance, Fiscal Reference Tables 2019, Table 17

Figure 19.10: Federal government public debt ratio, 1991–92 to 2017–18



Source: Department of Finance, Fiscal Reference Tables 2018, Table 15, Statistics Canada Table 36-10-0103-01 and author's calculations

In the years from 1996 to 2007 the Government of Canada had budget surpluses. Things were different in the years before 1995. Large budget deficits, averaging more than 5 percent of GDP, were the norm in the late 1980s and early 1990s. As a result, the outstanding federal government public debt increased, and increased faster than GDP, pushing the ratio of public debt to GDP up from 38 percent of GDP in 1983 to 68 percent of GDP in 1996. The cost of the interest payments the government had to make to the holders of its bonds increased from $3.9 billion to $42.4 billion. These costs accounted for almost 30 percent of budgetary expenses in 1995.

As a result, Canadian fiscal and budgetary policy shifted in 1995 to focus much more on deficit and debt control than on income stabilization. As Figure 19.9 shows, the federal budget was in surplus from 1997 until 2007. This reduced the debt to GDP ratio each year until 2008 as in Figure 19.10.

The financial crisis and recession of 2008-2009 and the federal government's *Economic Action Plan* changed this focus and pattern of fiscal policy. Fiscal stimulus through increased government expenditures and modest tax credits together with the recession in income created budget deficits in 2008-2013. These deficits added directly to the federal government debt. Larger debt combined with little or no growth in nominal GDP in 2009 and 2010 caused a sharp increase in the debt ratio shown in Figure 19.10. The focus of Federal Budget plans for recent years, as the recovery from recession seems to be underway, has shifted back to budget deficit control and reduction and a return to lower public debt ratios.

Although cumulative deficits can raise the public debt dramatically, it is not the absolute value of the outstanding debt that should be of interest or concern. If, at the same time the debt is rising, the economy is growing and tax revenues are rising as a result of a growing tax base, the government may be able to service a growing debt without having to raise taxes. The **public debt ratio (PD/Y)** is then the appropriate measure of the debt situation. A rise in the outstanding debt is not in itself a source of concern. However, the government cannot allow the debt ratio to rise without limit.

> **Public debt ratio (PD/Y)**: the ratio of outstanding government debt to GDP.

Recent sovereign debt crises in Portugal, Ireland, Greece and Spain provide clear examples of the difficulties high and rising public debt ratios cause. In those countries and others in Europe, and in the US, the government costs of rescuing banks in financial distress after 2008 combined in many cases with already large budget deficits, compounded by the recession in economic growth raised public debt ratios sharply. Table 19.2 showed the increase in public debt ratios from 2007 to 2015. The consequence was a loss of financial market confidence in the ability of some of these countries to pay interest on and subsequently retire outstanding government bonds, let alone service new bond issues to finance current deficits. Interest rates for new bond issues increased sharply and Greece and Ireland needed financial bailouts from joint EU rescue funds. This has provided time for fiscal adjustment but the economic growth required to solve sovereign debt issues remains elusive.

We will return to the relationship between fiscal policy and the public debt and examine the dynamics of the public debt ratio in Chapter 23.

This completes our introduction to the government budget, fiscal policy, aggregate expenditure, and the economy. It recognizes the importance of another set of linkages and feedback effects within the macro economy. We have seen two ways in which the government sector affects aggregate expenditure and output. Government expenditure is a part of autonomous aggregate expenditure. It affects the position of the AE function, equilibrium output, and the aggregate demand curve. The net tax rate diverts income from the private sector to government revenue and reduces induced expenditures. It changes the slope of the AE function, the size of the multiplier. Together the change in autonomous government expenditure and the effect of the net tax rate on induced expenditure change equilibrium income, and the AD curve. The size and direction of the change depends on the government's actual budget balance.

At the same time, conditions in the economy have significant effects on the government's budget. Recessions cause induced budget deficits and booms generate budget surpluses, both of which complicate the evaluation of the government's policy stance. Furthermore feedback effects within the government budget provide revenue support for fiscal expansion and revenue resistance to fiscal restraint or austerity. These effects mean government budget cannot be judged in the same way a private or household budget might be.

Federal government policies designed to suppress the COVID-19 pandemic, while at the same time moderating the impact of those polices on economic activity, called for a dramatic shift in budgetary policy. The multitude of programs created to give financial support for businesses and individuals in meant large increases in program expenses. At the same time budget revenues could be expected to decease and budget deficits would rise. Funding these deficits would require financing through bond sales and a corresponding increase in the federal public debt and the public debt to GDP ratio. Table 19.4 reports early estimates of the effects on the budget.

Table 19.4: Estimates of Government of Canada budgetary balances and public debt ratios, 2018-19 to 2020-21

	2018-19	**2019-20**	**2020-21**
		$ billions	
Total budget revenues	332.2	339.9	314.2
Total program expense	322.9	343.6	475.4
Public debt changes	23.2	23.7	23.0
Budgetary balance	−14.0	−27.4	−184.2
Federal debt	685.5	713.1	897.3
Federal debt/GDP, %	30.8	31.0	41.4

Source: Scenario Analysis Updates COVID-19 Pandemic and Oil Price Shocks, April 9, 2020. Revised PBO Report https://www.pbo-dpb.gc.ca/web/default/files/Documents/Reports/RP-2021-004-S/RP-2021-004-S_en.pdf. Appendix C.

19.7 Aggregate demand & equilibrium GDP

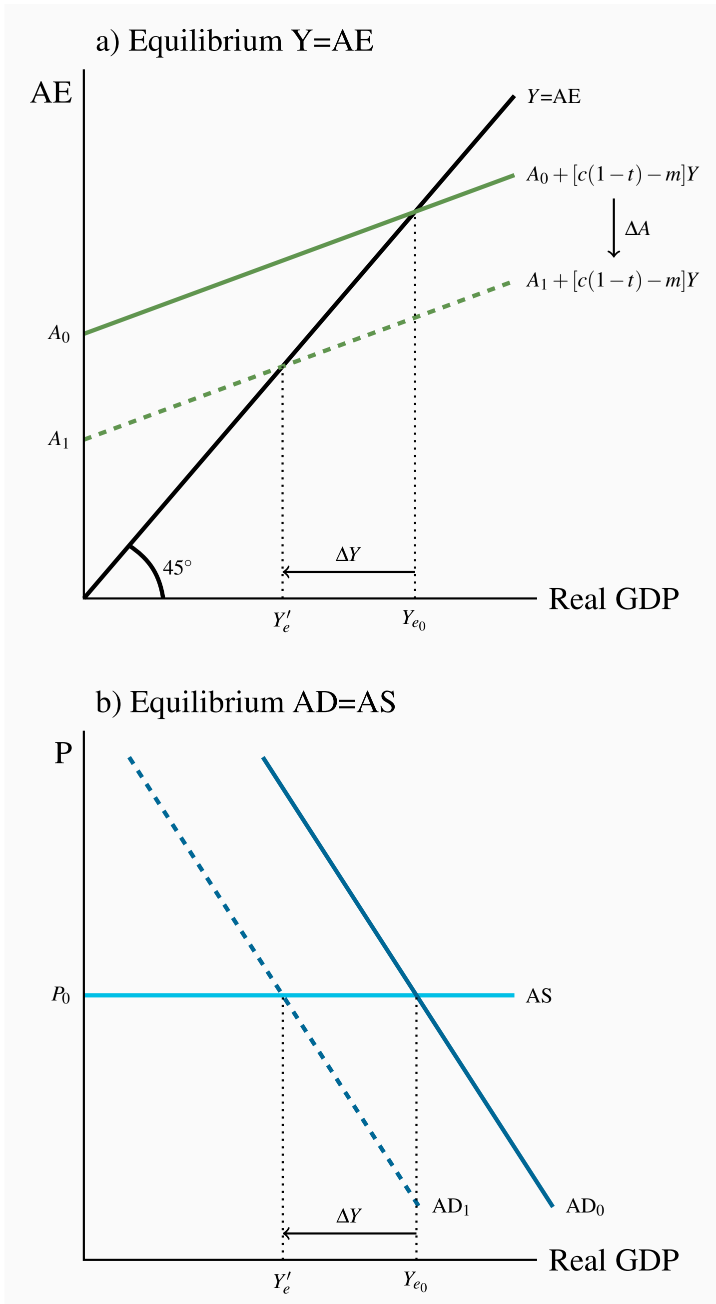
Our objective in this chapter was to extend the model of Chapter 18 to include a government sector and fiscal policy in aggregate demand. To do this we continued to assume that *wages, prices, money supply, interest rates, and foreign exchange rates are constant*. We also continued to make the important distinction between *autonomous* expenditure and *induced* expenditure, which leads to the existence of a multiplier. The equilibrium condition is still Y =AE and AD=AS, output and income equal to planned expenditure. Even though the model is more complex, it still shows us that fluctuations in autonomous expenditures, working through the multiplier, cause fluctuations in Aggregate Demand, output, income, and employment.

Changes in autonomous expenditures are still the sources of business cycles. If business changes planned investment expenditure in response to changed expectations about future markets, or if changes in economic conditions in other countries change exports or imports, the multiplier translates these changes into larger changes or fluctuations in income and employment. Government expenditure plans and net tax rates are fiscal policy tools that could be used to moderate or offset these fluctuations through a combination of automatic and discretionary fiscal policy.

Figure 19.11 shows the relationship between equilibrium income and output, and the link between changes in aggregate expenditure, aggregate demand, and equilibrium income. In the upper diagram a fall in autonomous expenditure from A_0 to A_1 reduces AE and equilibrium Y from Y_{e_0} to $Y_e^{'}$, which is the fall in A times the multiplier.

The fall in autonomous expenditure and equilibrium is a leftward shift in the AD curve in the lower diagram. The size of that shift in AD is the change in equilibrium income in the upper diagram, namely the fall in A times the multiplier. Because the price level is constant, giving a horizontal AS curve at P_0, the fall in equilibrium determined by AD/AS is the same as the horizontal shift in AD.

Figure 19.11: AE, AD and equilibrium output



Example Box 19.1: The effect of the government sector on equilibrium income

a) Equilibrium with no Government

Autonomous expenditure	$= A_0$	Autonomous expenditure	$= 80$
Induced expenditure	$= (c-m)Y$	Induced expenditure	$= (0.8-0.2)Y = 0.6Y$
Aggregate expenditure	$= A_0 + (c-m)Y$	Aggregate expenditure	$= 80 + 0.6Y$
Equilibrium income:	$Y = \text{AE}$	Equilibrium income:	$Y = \text{AE}$
	$Y = A_0 + (c-m)Y$		$Y = 80 + 0.6Y$
	$Y[1-(c-m)] = A_0$		$Y[1-0.6] = 80$
	$Y = A_0/[1-c+m]$		$Y = 80 \times 1/[1-0.6] = 80 \times 2.5$
			$Y = 200$

b) Equilibrium with added government sector: $G = 25, NT = 0.10Y$

Autonomous expenditure	$= A_0 + G_0$	Autonomous expenditure	$= 105$
Induced expenditure	$= c(1-t) - m$	Induced expenditure	$= [0.8(1-0.1)-0.2]Y = 0.52Y$
Aggregate expenditure	$= A_0 + G_0 + [c(1-t)-m]Y$	Aggregate expenditure	$= 105 + 0.52Y$
Equilibrium income:	$Y = \text{AE}$	Equilibrium income:	$Y = \text{AE}$
	$Y = A_0 + G_0 + [c(1-t)-m]Y$		$Y = 105 + 0.52Y$
	$Y[1-[c(1-t)-m]] = A_0 + G_0$		$Y(1-0.52) = 105$
	$Y = [A_0 + G_0]/[1-[c(1-t)-m]]$		$Y = 105/(1-0.52) = 105 \times 2.08$
			$Y = 218.4$

In this example adding government expenditure $G = 25$ financed in part by a net tax rate $t = 0.10$ raised autonomous expenditure by 25 but lowered the multiplier from 2.5 to 2.08 with the result that equilibrium income increased by 18.4. Equilibrium income increased because the government sector ran a budget deficit. G was 25 but net tax revenue was only $0.10 \times 218.4 = 21.84$.

Key Concepts

Canadian governments directly buy about 25 percent of GDP according to the national accounts data. They also spend about 17 percent on transfer payments to persons and business, including interest payments to holders of government bonds.

Government expenditure G on goods and services, including the public services provided to households and business is a policy variable and an autonomous component of aggregate expenditure.

Net taxes $(NT = tY)$, the revenue collected by government from households, are difference between taxes collected and transfers paid.

Disposable income is national income minus net taxes. Changes in disposable income cause changes in household consumption expenditure based on the MPC.

The **net tax rate (**t**)** reduces changes in disposable income relative to national income and reduces the marginal propensity to consume out of national income to $c(1-t)$. This lowers the slope of AE and the size of the multiplier.

Government expenditure and net taxes affect equilibrium national income by changing both autonomous expenditure and the multiplier.

The **government budget** describes what goods and services the government plans to buy during the coming year, what transfer payments it will make, and how it will pay for them. Most spending is financed by taxes, but some revenue comes from charges for services.

The **government budget balance** is the difference between net revenues and government expenditures. Because net tax revenues depend on national income $(NT = tY)$ the **actual budget balance** is determined by the government's budget plan and the level of national income. The actual budget balance will change with changes in national income.

A **balanced budget** has revenues equal to expenditures.

A **budget surplus** means revenues are greater than expenditures.

A **budget deficit** means revenues are less than expenditures.

Fiscal policy is the government's use of its taxing and spending powers to affect aggregate demand and equilibrium GDP.

The **structural budget balance** (SBB) is an estimate of what the budget balance would be if the economy were operating at potential output. *Changes in the structural budget balance* are indicators of changes in fiscal policy because they measure changes in expenditure and tax programs at a standardized income level.

The government's tax and transfer programs are **automatic (fiscal) stabilizers** that reduce the size of the multiplier and the effects of *transitory* fluctuations in autonomous expenditures on equilibrium GDP.

Discretionary fiscal policies are changes in net tax rates and government expenditures intended to offset *persistent* autonomous expenditure shocks and stabilize aggregate demand and equilibrium output at potential output.

Public debt (PD) is the outstanding stock of government bonds issued to finance past government budget deficits minus the retirement of government bonds in times of past government budget surpluses. The annual change in the public debt is $\Delta PB = -BB$.

Public debt ratio (PD/Y) is the ratio of outstanding government debt to GDP, PD/Y.

Recent **sovereign debt crises** in Portugal, Ireland, Greece and Spain provide clear examples of the difficulties high and rising public debt ratios cause.

The **government sector and fiscal policy** are important determinates of aggregate demand and equilibrium GDP. Government expenditures are an autonomous policy variable. Net tax rates and policy affect the size of the multiplier. Changes in government expenditure and tax programs through the setting of the government's budget affect AE, AD and equilibrium GDP.

Exercises for Chapter 19

Exercise 19.1 Suppose a government is established in a country where none previously existed. The government spends 100, financed by borrowing, to provide public services. If autonomous expenditure before government is set up is 200 and induced expenditure is $0.6Y$ (based on $mpc = 0.75$ and $mpm = 0.15$) what is the equilibrium value of real GDP before the government is established and what would equilibrium GDP be after government is established?

Exercise 19.2 If the government expenditure in Exercise 19.1 were financed by imposing a net tax rate on income of $t = 0.10$:

(a) Calculate and compare the slopes of the AE functions in Exercises 19.1 and 19.2.

(b) Calculate and compare the multipliers in Exercises 19.1 and 19.2.

(c) What is the equilibrium real GDP in Exercise 19.2 compared to Exercise 19.1?

Exercise 19.3 If government expenditure is 100 and the net tax rate is $t = 0.20$:

(a) Complete the following table:

Y	$NT = tY$	G	$BB = NT - G$
100			
200			
300			
400			
500			
600			
700			

(b) In a diagram with national income Y on the horizontal axis and government revenue and expenditure on the vertical axis, draw the government expenditure and net tax functions. Explain the intercept on the vertical axis, and the slope you have given to the NT and G functions in your diagram.

(c) Suppose the government cuts the tax rate to $t = 0.15$. Show the effects in your diagram.

Exercise 19.4 Using diagrams illustrate an initial equilibrium national income, the effect of the increase in government expenditure on that equilibrium national income, and the government's budget functions and balances before and after the increase in government expenditure.

Exercise 19.5 Suppose the government raises its revenue by a net tax of 25 percent on income, $t = 0.25$. Further suppose that induced expenditure is $0.45Y$ (based on $c = 0.8$, $t = 0.25$ and $m = 0.15$).

(a) What is the slope of the AE function? What is the size of the multiplier?

(b) Autonomous expenditure by the non-government sectors (A_0) is 300 and government expenditure is 400. What is the equilibrium income and output? What is the government's budget balance?

(c) Now assume the government increases its expenditures by 100 to provide additional funding for national defense. What is the effect on equilibrium income and output? What is the effect on the government's budget balance?

Exercise 19.6 An economy is in equilibrium at a real GDP of 750, but current estimates put potential output at $Y_P = 850$.

(a) Is there an inflationary or a recessionary gap, and, if there is either, what is its size?

(b) Research suggests that induced expenditure is $0.5Y$ (based on $c = 0.75$, $m = 0.10$, and $t = 0.20$). If there is a gap, what change in government expenditure would eliminate the gap?

(c) If the government preferred to change its net tax rate to eliminate the gap, and not change government expenditure, what new tax rate would be required to eliminate the gap?

Exercise 19.7

(a) Draw a diagram that shows the government's budget balance relative to national income. Explain briefly the vertical intercept of the budget function and its slope.

(b) Using your diagram from (a), show the structural budget balance and a situation in which the actual balance is different from the structural balance.

(c) Based on this diagram, show and explain the difference between the budget effects of automatic stabilization and discretionary fiscal policy.

Exercise 19.8 Suppose as in Exercise 19.5 the government raises its revenue by a net tax of 25 percent on income, $t = 0.25$. Further suppose that induced expenditure is $0.45Y$ (based on $c = 0.8$, $t = 0.25$ and $m = 0.15$), autonomous private expenditure is 300 and government expenditure is 400.

(a) What is the government's budget balance?

(b) If the government has outstanding public debt equal to 500 what is the public debt to GDP ratio?

(c) Suppose the government increases its expenditure by 100 without any increase in the tax rate. What are the new equilibrium GDP and the new budget balance? Explain why the change in the government's budget balance is different than the change in government expenditure.

(d) What is the outstanding public debt and the public debt ratio in the new equilibrium, assuming the economy has reached its new equilibrium national income in one year?

Part Eight

Financial Markets & Economic Activity

Three chapters in this part examine the financial sector of the economy and the important role it has in the determination of output and employment. They cover money, banks and banking, financial markets, asset prices and foreign exchange rates. When integrated, these parts of the economy provide the 'monetary transmission mechanism' and a framework for the design and implementation of central bank monetary policy. Moreover, the financial sector is the key to price level and inflation rate elasticities in the aggregate demand functions derived in later chapters.

Chapter 20

Money, banking & money supply

In this chapter we will explore:

Today banks are still the front line of the financial system. Bank deposits are an important form in which people hold their wealth and they are also a means of payment, making them the main component of the money supply in the economy. Furthermore banks are the major sources of credit for households and businesses, either directly through bank loans or indirectly through the credit card systems they own and third party systems they service. Banks are the foundation of the payments system.

But the financial and payments system has changed a lot in the past ten years and continues to evolve. Now *fintech* is seen as the new way for handling payments, insurance, investing and lending. *Cryptocurrencies* and crypto-assets such as Bitcoin, Ethereum and Libra, and *blockchain* networks they use for accounting and payments are major topics for discussion and analysis. Still the Bank of Canada does not see any of the crypto-assets as money right now. The banks and the banking system continue as the third parties necessary for the basic functioning of the monetary system.

This chapter examines the evolution of the financial sector and the important link it makes between aggregate expenditure and aggregate demand. *It is the key to understanding the slope and position of the AD curve*. It also creates the framework for central bank *monetary policy*.

20.1 Money and the functions of money

You can see the variety of things that have been used as money in Canada in James Powell's *A History of the Canadian Dollar*, available at the Bank of Canada's website: `https://www.bankofcanada.ca/2005/12/a-history-of-the-canadian-dollar-by-james-powell/`.

Our money in the seventeenth and eighteenth centuries was silver and gold coins from many countries, and playing cards. The British pound sterling, the Spanish silver dollar, and the US dollar were the main moneys in Canada in the nineteenth century, followed by paper currencies issued by banks and by the government since the late nineteenth century.

It is not the commodity or token used as money that matters, but the social convention that it is trusted and accepted without question as a means of payment. Money makes it easier for everyone to buy and sell goods and services and economize on the use of scarce resources.

Money is defined by four important functions. It provides:

1. A means of payment as a medium of exchange
2. A unit of account
3. A store of value
4. A standard of deferred payments

As a **means of payment** money is involved in most exchanges. We use money to pay for the goods and services – from food and clothing to transportation, to rent, to fees and taxes. People receive wages and salaries and other types of income in money. Money is not consumed in these transactions. It is used as a medium of exchange.

> **Means of payment**: a commodity, token, financial asset, or transferable electronic asset such as a bank deposit, generally accepted in payment for goods and services or the repayment of debt.

Exchange transactions without money are **barter exchanges**, a direct exchange of one good for another. These exchanges depend on a *double coincidence of wants*. Each party to the exchange has a good the other party wants and is willing to trade one good for another. This means exchange transactions are time consuming and expensive as people must find others who have what they want and want what they have. Using a money as a medium of exchange dramatically lowers the cost and increases the ease and efficiency of trade.

> **Barter exchanges**: direct exchanges of goods or services for goods or services without the use of money.

Money also serves as a **unit of account**. Prices in Canada are quoted in Canadian dollars. Similarly in other countries prices are quoted in domestic currency. In much of Europe prices are in euros, in the United States in US dollars and in Japan in yen. This reflects the convenience of using the same units for the means of payment and the unit of account. However, there are exceptions. Historically, in Canada, during the time of the fur trade, books were kept in "currency" but actual currency never changed hands in the barter of trade goods for furs.

Unit of account: the standard in which prices are quoted and accounts are kept.

To serve as a medium of exchange, money *must also be* a **store of value**. Money works as a time machine allowing people to separate the timing of their expenditures from the timing of their incomes. Money is accepted today with *confidence* that it can be used some time in the future to make payments when buying goods and services. You would not accept money today that you thought would be unacceptable when you offered it in payment at some later date. The integrity of the issuer of money and trust in the institutions that hold your money balances are essential.

Store of value: an asset that carries purchasing power forward in time for future purchases.

Money is not a unique store of value. Other assets including real estate, financial assets like corporate and government bonds, fine art and antiques all serve as stores of value. These may be better ways to store value, but people still choose to hold some of their wealth as money. This choice to hold money balances is very important to the effects money balances have on financial markets and aggregate expenditure. Chapter 21 examines it in detail.

Money provides a **standard for deferred payments**. If you take out a student loan the amounts you will repay in the future are specified in dollars. Similarly, servicing and retiring a mortgage on a property or a loan on a car calls for future payments specified in dollars. Domestic money is not essential for this function. Individuals, businesses and governments often borrow or lend in the money of other countries. In those cases the currency in which the loan transaction takes place is usually the standard for payments to settle the debt. The essential attribute of money is its general acceptance as a means of payment. For this money must also be a store of value. This works well when money is also a unit of account and a standard of deferred payments.

Standard of deferred payments: the units in which future financial obligations are measured.

The development of money

The money we use today is the product of a long and continuing evolution in the financial services industry. It is a testament to the ingenuity of people and society seeking to reduce the costs and increase the volume of trade in goods and services.

Historically, there were no banks. Money was a commodity. Gold and silver bullion are two commodities that came to be used extensively because of their relative scarcity and durability. Concerns about the purity of these metals and the inconvenience of weighing out small quantities to make payments led to coinage. The minting of gold and silver coins by heads of state offered a solution to these problems. The 'monarch' certified the purity and quantity of the metal in the coin by having his or her likeness stamped into the metal.

Unfortunately, coinage did not completely solve the concerns about the quantity and quality of gold and silver money. The integrity of the coinage process was still uncertain. The quantity of gold in a coin could be reduced by clipping its edges, or by rubbing the surfaces of the coin to wear some of the metal away. "Sweating" coins by placing them in a leather bag and shaking them was one technique used to remove some of their precious metal. The edge designs, millings, or facets that we still see on coins today were introduced to combat clipping, and wear on the heads and tails stamped into coins provided evidence of sweating. Coins that were worn or clipped were not accepted at full value in exchange for goods and services.

A second difficulty with precious metal coinage came from the sovereign who controlled the minting process. Adding a little base metal to the gold or silver being minted resulted in coins with less precious metal content than their face value certified. A little of the precious metal was withheld and could be used to mint more coin, which was, in effect, free money for the sovereign. This "debasing" of the coinage was a serious problem at times and, like clipping and sweating, reduced the acceptability of precious metal coinage as money.

The emergence of banks and paper money was a response to the problems with gold and silver commodity money. The first banks were goldsmiths who used gold in the production of jewelry and ornaments. They were accustomed to buying and selling gold bullion, and holding stocks of gold bullion. It was a natural extension of their business to offer to hold gold for safekeeping. Those who stored their gold with goldsmiths were given paper claims or receipts (IOUs), which were redeemable in gold on demand. Confidence in redemption at face value was the key to acceptability.

When people began to use gold receipts to make payments, gold receipts became a means of payment. They were **token money**, representing a fixed amount of the precious metal.

> **Token money**: redeemable claims on commodity money.

Goldsmiths became bankers when they realized that not all their customers would show up at the same time and want their gold back. Trust in the ability to redeem gold receipts made them acceptable as a medium of exchange. Gold merchants could make loans by issuing more gold receipts than they had gold in their storage vaults. They only needed gold holdings equal to a fraction of the gold receipts they had issued, as long as people used the receipts as a medium of exchange.

Banks as we know them grew out of this acceptance by society of credit (IOU) money as a medium of exchange. Banks began to accept customer deposits of token money and to issue their own bank notes (credits) as receipts. People liked the convenience and safety of storing some of their wealth with banks. As society became more comfortable with banks and confident in the safety of banks, bank deposits, which could be transferred by cheque, became widely accepted as the medium of exchange. Bank notes and deposits were no longer redeemable in gold or commodity money, but they were redeemable in **legal tender**. Governments established central banks to control the supply of legal tender, bank notes, or cash. Bank notes now serve as both a medium of exchange and as the **reserves** banks hold to ensure the redemption of their customers' deposits on demand.

> **Legal tender**: money that by law must be accepted as a means of payment.
>
> **Bank reserves**: cash (legal tender) held by banks to meet possible withdrawals by depositors.

Unlike other financial institutions, such as pension funds and insurance companies, the key aspect of banks is that some of their liabilities are used as the medium of exchange; cheques and debit cards allow their deposits to be used as money to make payments. Bank deposits are credit money.

In Canada today, as in most industrial countries, we use a combination of **fiat money** and **credit money**. Fiat money, in contrast to commodity or token money is money that the government has declared to be legal tender. Coins and paper bank notes are fiat money in Canada. If you look carefully at a \$5, \$10, or \$20 Bank of Canada bill you will find the statement: "This note is legal tender." By law it must be accepted as a means of payment for goods and services bought or debts repaid.

> **Fiat money**: money the government has declared as legal tender.
>
> **Credit money**: the debt of a private business or individual.

Our fiat money is supplemented by credit money. A bank deposit is credit money, and is redeemable in fiat money on demand, or in the case of some savings and time deposits, after a short waiting period. Furthermore, the bank is obliged to pay when your cheque is presented, or when you use your debit card. Bank deposits are a medium of exchange because they are generally accepted as a means of payment, even though they are not legal tender. The sum of bank deposits and fiat money in circulation outside the banks at any time is the stock of medium of exchange and the economy's **money supply**.

> **Money supply**: the means of payment in the economy, namely currency (notes and coin) in circulation outside the banks and bank deposits.

Payment methods, monies and financial services continue to evolve as illustrated by **Fintech**, the ongoing development of **e-payments** and **e-money** and more broadly.

> **Fintech**: the rapidly growing, innovative, applications of computer and internet technologies to facilitate the flow of payments and financial information.

E-payments are now familiar and widely used. Debit cards, telephone banking, internet banking, e-transfers, wireless and mobile banking provide access to manage deposit accounts and make payments, but they do not involve any creation of money. The central bank still controls money supply.

E-monies fall into two categories. **Centralized e-money**, for example a multi-purpose pre-paid payment card, denominated a national currency, has money value based on funds received by the issuer. It is accepted as a means of payment in transactions with parties other than the issuer. Stored-value cards and gift cards for purchases at specific companies use the Visa or MasterCard systems. They are widely accepted in payment. In Canada, transactions using these and similar cards are denominated in Canadian dollars.

> **Centralized e-money**: a payment card denominated in a national currency.

Cryptocurrencies are a form of decentralized e-money. They are financial assets but have no centralized issuer and are not denominated in any national currency. The *Bitcoin* and *Ethereum* are examples. Each is completely decentralized and does not represent a claim on the issuer in the way that a Canadian dollar bank deposit represents a claim on a bank. Furthermore, no bank or institution is an intermediary in a transaction between payer and payee. A cryptocurrency is decentralized over a peer-to peer computer network that directly links users, but no one user is in control of the network. It uses 'blockchain technology' which is a shared, continually reconciled data base. However, because cryptocurrencies are not redeemable in national currencies, and are not legal tender, they generally trade online and values are highly volatile in terms of national currencies. They are not money in that they do not serve as a *'generally acceptable'* means of payment.

> **Cryptocurrency**: a decentralized e-money with no centralized issuer and not denominated in any national currency.

The recent *Facebook* proposal for a cryptocurrency, *Libra*, and a *Calibra* wallet, try to make some innovations to overcome volatility, convertibility and risk concerns raised by recent experiences with cryptocurrencies. For one, the proposed currency and transactions are to be entrusted to a *'Libra Association'*, a membership organization funded by *Calibra* and range of twenty-seven other businesses and associations. The apparent objective is to give confidence in, and assurance about, the integrity of operations. Additionally, to address market value volatility and convertibility concerns raised by experience with *Bitcoin* and others, the value of the 'Libra token' is to be fully backed by financial assets such as a 'basket of national currencies'. But this is not a fixed basket and the mix of currencies might change, defeating the intention to provide stability in the trading price of the *Libra*. Even these innovations do not appear to satisfy the trust, security and volatility issues.

For example concerns arise about the membership and operations of the Libra Association. The structure membership and operations of the 'Association' might not be stable. Also the financial assets the 'Association' would manage would be purchased with user funds, but earnings on these funds would not flow to user wallets. What are the investment criteria and objectives? More broadly, there are important reservations about giving a social media company that has not demonstrated trustworthiness, access to personal financial information.

These are interesting developments but in the opinion of the Bank of Canada none of the current crypto assets is money but may be useful financial portfolio assets. The need for trust in third party intermediaries like banks remains. Facebook's Libra proposal is unlikely to change this.

20.2 Measures of the Canadian money supply

The **money supply** is traditionally defined as cash in circulation outside the banks, plus bank deposits. But as the banking and financial system evolved so did the types of deposits issued to the non-bank public. Now there are questions about the measurement of money supply.

> **Money supply**: notes and coin in circulation outside banks plus bank deposits.

In the early days of banking there were demand deposits on which cheques could be written and savings deposits which often required a period of notice before funds could be withdrawn. Today banks offer a much wider spectrum of deposits to customers from demand to savings deposits that may or may not be chequable, pay interest under different terms, and some which can only be accessed online. Not all deposits serve as means of payment. For these the balance must be transferred to another account before it is available to make a payment. Which deposits should be counted in the money supply?

The structural evolution of the financial system raises further questions. What is a 'bank'? Today banks compete vigorously for deposits with other businesses, including trust companies and credit unions whose deposits are widely accepted as means of payment. There is no longer a reason to exclude those deposits from measures of the money supply. Different measures of the money supply illustrate the importance of different financial institutions in the industry.

The Bank of Canada now publishes data on the **monetary base** in response to continuing changes in the types of bank deposits available to households and businesses. Advances in technology, financial deregulation, and competition in the financial services sector, which have led to more types of financial assets, make it easy for customers to substitute between those assets we include in narrow definitions of money supply and other assets. But once we leave the monetary base as a measure of money supply, there is no single measure of money that is clearly the means of payment. There is, however, only one type of money that is legal tender; namely, notes and coin.

> **Monetary base**: legal tender comprising notes and coins in circulation plus the cash held by the banks.

Table 20.1 shows the size of the money supply in Canada based on different definitions and money aggregates. These data illustrate the range of choice involved in the selection of a specific measure of the money supply. But one thing is clear: Bank deposits are the major component of money supply by any measure other than the currency component of the monetary base.

Table 20.1: The money supply in Canada in September 2018 (billions $)

Monetary base (MB)	94.2
Currency in circulation (CR)	86.9
M1B = currency + chequable chartered bank deposits	878.0
M2 = M1B + notice and savings deposits in the banks	1,632.6
M2+ = M2 + deposits at other financial institutions	2,058.6

Source: Statistics Canada, Table 10-10-0166-01

Currency in circulation is only about 6 percent of M2. Deposits account for the remaining 94 percent. The importance of bank deposits as money means that understanding the operations of banks as sources of loans and deposits is the key to understanding the money supply function in the economy.

20.3 Banking in Canada today

In Canada today, and in other industrial countries, the mainstream banking system is made up of a *central bank* and a number of *commercial banks* and other deposit-taking institutions called *near banks*. Table 20.2 illustrates the structure of the banking industry in Canada. The industry is defined broadly to include deposit-taking institutions, not just those that operate under the federal Bank Act.

Table 20.2: The Canadian banking system, June 2018

Banking Institution	**Number**
Central Bank:	
The Bank of Canada	1
Number of Chartered Banks:	
Schedule I domestic banks	30
Schedule II foreign banks subsidiaries	17
Schedule III foreign bank branches	22
Total	**70**

Source: Canadian Bankers' Association. http://www.cba.ca/memberbanks

Banks are **financial intermediaries**. They borrow money from the public, crediting them with a deposit. The deposit is a liability of the bank. It is money owed to depositors who are confident it will be redeemed on demand. The money raised from depositors provides the funds to support the bank loans made to businesses, households, and governments.

> **Financial intermediary**: a business that specializes in bringing borrowers and lenders together.

Banks are not the only financial intermediaries. Trust companies, credit unions, caisses populaires, insurance companies, securities dealers, mutual fund companies, and independent financial advisors all play a role in this industry. But banks hold more than 70 percent of the assets in the financial services sector, and the six largest Canadian banks account for over 90 percent of the assets of the banking industry. Trust companies, credit unions, and caisses populaires also accept deposits that are used as money, but those deposits are a small fraction of the total of deposit money. As a result, bank deposits are the focus of our discussion of money in Canada.

The **Bank of Canada** is Canada's central bank. It is the source of the bank notes used to make payments and held as cash reserves by commercial banks. Established by the government in 1935, it has the responsibility to regulate the national money supply and support the operation of financial markets. The Bank's power to meet these responsibilities comes from its monopoly on the issuance of bank notes.

> **Bank of Canada**: Canada's central bank.

The Bank of Canada also is the provider of:

- Banking services for the commercial banks in the system
- Banking services for the federal government
- Lender-of-last-resort facilities in times of liquidity crises and reserve shortfalls

Commercial banks hold some of their reserves as deposits in the Bank of Canada, and make payments among themselves using their Bank of Canada deposits. These interbank payments arise from wire transfers, direct deposits, pre-authorized debits, bill payments, point-of-sale debits, and online payments made by bank customers. For example, cheques written by customers at one bank, say *Scotiabank*, but paid to and deposited by customers of the *Royal Bank* result in transfers of deposits between these banks. To settle these transfers, *Scotiabank* must pay the Royal Bank. Funds held by *Scotiabank* on deposit in the Bank of Canada are used for this purpose. They are called "settlement balances." In 2018, *Payments Canada* (`www.payments.ca`), which coordinates this clearing of interbank transactions, handled more than 21.9 billion transactions including cheques, credit card and e-transfers and point-of-sale debits totalling $97 trillion.

The government holds some deposits in the Bank of Canada. Government receipts, like income taxes paid to the Receiver General, are deposited in government accounts in the Bank of Canada.

Government payments like Old Age Security, Employment Insurance benefits, bond interest, and income tax refunds are paid with government cheques or transfers drawn on its Bank of Canada account. Government funds over and above those needed to make regular payments are held on deposit in the commercial banks, and earn interest income for the government.

The key difference between a central bank and the commercial banks in the banking system is the profit motive. Central banks *do not* pursue profits. Their operations focus on the management of the cash reserves available to the public and the banks. The supply of cash reserves affects the behaviour of other banks and financial markets more generally. This is the monetary policy role of the central bank. We will examine it in detail in Chapter 22.

Commercial banks, on the other hand, are profit-oriented businesses. They operate, as we will see shortly, to maximize the profit they earn for their owners. To this end, they offer banking services to the public. Using the notes and deposits issued by the Bank of Canada as reserves, they issue bank deposits to their customers—which are widely used as the medium of exchange—and they make loans to finance purchases made by businesses and households.

To illustrate the business of these banks, Table 20.3 shows the consolidated balance sheet of Canadian chartered banks in April 2019. In the table we see that the banks held small cash balances as reserves against their deposit liabilities. Their other Canadian assets were mainly loans to households and businesses, including mortgage loans, and their holdings of financial securities. Because cash and many of their financial securities have high **liquidity**, banks can make long-term loans and still have cash and funds available if depositors withdraw their money.

> **Liquidity**: the cost, speed, and certainty with which asset values can be converted into cash.

Table 20.3: Balance sheet of Canadian chartered banks, April 2019

Assets	**billions $**	**Liabilities**	**billions $**
Canadian dollars:		Canadian dollars:	
Cash	42.6	Personal deposits	1,038.8
Government of Canada securities	285.6	Non-personal deposits	916.0
Corporate securities	171.7	Government deposits	24.0
Personal and business loans	1,115.1	Advances from Bank of Canada	33.0
Mortgages	1,172.0	Other liabilities	413.8
Foreign currency assets	3,102.0	Total Cdn dollar liab & shareholder equity	2,874.0
Other assets	62.3	Total for currency liab & shareholder equity	3,077.3
Total assets	**5,951.3**	**Total liabilities and shareholders' equity**	**5,951.3**

Source: Bank of Canada, Banking and Financial Statistics, April 2019: *Chartered bank assets: Month end*, and *Chartered bank liabilities and Shareholders equity*. Formerly Tables C3 and C4 and author's calculations. Figures have been rounded to one decimal place.

However, many loans to businesses and households are quite *illiquid*. The bank cannot easily get its money back in a hurry. This is not really a cause for concern when people and businesses have confidence in the banks and make widespread use of bank deposits as money. Payments and receipts are both in bank deposit form, which are cleared quickly and efficiently through the cheque-clearing and transfer facilities. Banks need only small cash balances to cover the net clearings and net public demand for cash. In Table 20.3, the banks are holding only $42.6 billion against deposit liabilities of $1,978.8 billion.

Canadian banks also carry on important international banking operations, as do banks in many other countries. We see this business recorded on the balance sheet as foreign currency assets and liabilities. The foreign currency assets are mainly loans to customers and holdings of foreign financial securities. Foreign currency deposits of customers are the main foreign currency liabilities. These foreign currency operations are similar to the banks' domestic currency operations. The banks provide loan financing to customers needing foreign currency to make payments in other countries, and they provide deposit facilities for customers using foreign currency for international transactions.

Competition and co-operation are important to the efficient operation of the banking system. Banks compete among themselves for customer deposits and customer loans. Some of the competition for deposits is based on the location, convenience, and quality of bank branches, some on the offers of service packages including personal financial advice and wealth management, and some on the interest rates offered on deposit balances. If you watch media advertising, you are probably aware that some small banks like *Simplii Financial* and *Tangerine Bank* offer you a relatively high interest rate and will make no service charges if you put some of your funds on deposit with them. Success in attracting deposits is very important to size and growth of a bank's business.

Credit-worthy customers willing to borrow funds are equally important to a bank's operations. Interest income earned on customer loans is the major source of bank revenue. As a result, banks compete in the personal and business loan markets, using both the terms of loans and the interest rates charged on loans to attract borrowers. The market for mortgage funds is one of the most competitive areas of bank operations. Mortgage rates and terms are advertised widely in the media and in displays in bank offices and even in supermarkets.

Despite this competition for deposits and loans, the banking system depends on the co-operation among banks that makes deposits the medium of exchange. Co-operation in the cheque-clearing system and the debit card *Interac* system are two important examples of banks working jointly to provide the payments system. A cheque book or a debit card is not very useful if it can make payments only to other people or businesses that do business with the same bank you use. Joint interests in *VISA* and *MASTERCARD* are a second important part of inter-bank co-operation that makes these cards widely acceptable as a source of credit.

There are also important areas of bank co-operation on the lending side of their operations. It often happens that businesses and industries have projects that need more financing than any one bank can or wants to provide. However, several banks might agree to provide funding jointly, increasing their lending capacity and spreading the risks associated with the project among them.

These dimensions of competition and co-operation among banks, and their contribution to the efficient functioning of the money and financial sector of the economy, appear whenever questions arise about bank mergers in Canada.

Banking operations and profits

A commercial bank is a profit-oriented business. Its profits come from the difference between what it costs it to raise funds and the revenues it earns from lending. To bring deposits in, the bank offers customers a range of banking services, including safekeeping, record keeping, access to banking offices or bank machines, chequing, internet banking, e-transfers and debit card facilities, and interest income on some types of deposits. Service charges or fees cover the costs of some of these services. The interest payments to depositors are the main net cost of funds to the bank.

To be profitable, banks have to find ways to lend, at acceptable levels of risk, the funds they have attracted from depositors. Table 20.3 shows how banks lend their money. In Canadian dollars, most is lent to households and businesses at interest rates established for different types of personal, business, and mortgage lending. Some is used to buy government securities and other financial assets, usually with a short time to maturity. These assets pay a lower rate of interest than loans, but they are more liquid and provide the banks with funds if people decide to withdraw a lot of money from their deposit accounts. Notice that the banks also hold some cash, on which no interest is earned, to cover the day-to-day **clearing balances** that come from the withdrawals, deposits, and transfers made by their customers.

> **Clearing balances**: cash balances used by a bank to settle the differences between its customers' payments and receipts involving other banks.

Bank profits come from the difference or spread between the interest cost of raising funds from depositors and the interest income earned on bank assets. If, for example, the banks pay, on average, 4 percent on their deposit liabilities of all types and earn, on average, 6 percent on their assets of all types, their **net interest income** would be 2 percent. To take an actual example, the *Scotiabank Annual Report* for 2018 reports net interest income of 2.26 percent of average assets in 2018. *Scotiabank* net interest income was higher than the previous year as a result of asset growth in residential mortgages, consumer auto and commercial lending, which increased deposit to lending interest rate spreads. The other large banks report net interest income of the same order of magnitude but there are variations among them. The key to profitability is choosing the right mix of high-quality (low-risk) loans and investments while at the same time controlling the costs of raising funds.

> **Net interest income**: the excess of loan interest earned over deposit interest paid.

As we saw in Table 20.3 Canadian banks held only \$40.2 billion in cash against \$1,868.5 billion in personal and non-personal deposit liabilities. Their cash reserve assets were about 2.2 percent of their total deposits. The skill in running a bank entails being able to judge just how much must be held in liquid assets, including cash, and how much can be lent out in less liquid forms that earn higher interest income. The profit motive pushes the bank toward riskier, higher interest paying assets and higher net interest income. **Banker's risk**, the risk that customers will withdraw their deposits and demand cash, pushes the bank toward holding higher cash balances. But cash balances earn no interest income and reduce the bank's net interest income.

> **Bankers risk**: the risk that customers may demand cash for their deposits.

20.4 Money created by banks

Banks *create money* when they increase their *deposit liabilities* to pay for the loans they make to customers, or for the financial securities they buy. The public uses the deposit liabilities of the banks as money to make payments or to hold as a store of wealth. There are *four key conditions* that give banks the ability to create money:

1. The non-bank public has confidence in banks and is willing to hold and use bank deposits as money.
2. The non-bank public is willing to borrow from the banks to finance expenditure or asset purchases.

3. The banks are willing to operate with reserves assets, cash for net cheque clearings and withdrawals, equal to some small fraction of their deposit liabilities.
4. The banks are willing to accept the risks involved in lending to the non-bank public.

If any of these is absent, the banks cannot create money, although they may provide safekeeping services.

The first two conditions underlie the demand for banking services. Banks acquire cash by providing customers deposit services and bank customers use bank loans as a source for funds to pay for purchases of goods, services and financial assets like equities and bonds. If the non-bank public is unwilling to use bank services there is no banking industry.

The third condition required for the banks to create money is a bank reserve ratio that is less than one. The **reserve ratio (rr)** is the ratio of cash on hand to deposit liabilities that banks choose to hold.

$$rr = \frac{\text{reserve assets}}{\text{deposit liabilities}} \tag{20.1}$$

Reserve ratio (rr): the ratio of cash reserves to deposit liabilities held by banks.

Cash holdings are reserve assets. If banks choose to hold reserves equal to their deposit liabilities, $rr = 1$ and the banks cannot create deposits. They are simple safety deposit boxes.

A simplified case shows how banks can and do create deposits. Assume banks use a reserve ratio of 10 percent ($rr = 0.10$). Suppose initially the non-bank public has wealth of $1,000 held in cash, before they decide to switch to bank deposit money. This cash is a private sector asset. It is a liability of the central bank or government, which issued it, but not a liability of the private banks. The 'Initial position' in Table 20.4 uses a simple balance sheet to show this cash as an asset of the non-bank private sector.

Table 20.4: How the banking system creates money

Banks				Non-bank public			
Assets		Liabilities		Assets		Liabilities	
Cash	0	Deposits	0	Cash	1,000	Bank loans	0
Cash	1,000	Deposits	1,000	Cash	0	Bank loans	0
				Deposits	1,000		
Cash	1,000	Deposits	10,000	Cash	0	Bank loans	9,000
Loans	9,000			Deposits	10,000		

Then in the second part of the table people deposit this \$1,000 of cash into the banks by opening bank accounts. Banks get assets of \$1,000 in cash, distributed among individual banks by their customers and issue total deposit liabilities of \$1,000. These deposits are money the banks owe to their depositors. If banks were simply safety deposit boxes or storerooms, they would hold cash assets equal to their deposit liabilities. Their reserve ratio would be 100 percent of deposits, making $rr = 1.0$. Table 20.4 would end with part 2.

However, if the public uses bank deposits as money, the banks don't need all deposits to be fully covered by cash reserves. It is unlikely that all depositors will show up at the same time and demand cash for their deposits. Recognizing this, the banks decide that reserves equal to 10 percent $(rr = 0.10)$ of deposits will cover all *net* customer demands for cash. In this case, the banks have excess reserves which in total equal 90 percent of their deposit liabilities or, initially, \$900.

The banks use their excess reserves to expand their lending. Each bank makes new loans equal to its excess reserves. It pays for those loans by *creating an equal amount of deposits*. If you were to borrow from bank your personal deposit would be increased by the amount of the loan. The same thing happens to other people who borrow from their banks.

In our example, all banks combined can create \$9,000 of loans based on \$1,000 in new cash reserves. In part 3 of Table 20.4, we see loans of \$9,000, as assets on the banks' balance sheets, and \$9,000 of new deposits to customers, against which they can write cheques or make payments online or by transfers. The newly created deposits of \$9,000 are a part of the \$10,000 liability on the banks' balance sheets. The public now has bank deposit assets of \$10,000 and liabilities, loans owed to the banks, of \$9,000. Non-bank public net worth, assets minus liabilities is \$1,000, the cash they originally deposited in the banks. Because the public uses bank deposits as money, the banks can buy new loans by creating new deposits.

The reserve ratio is 10 percent in part 3 of Table 20.4 ($rr = \$1,000$ cash/\$10,000 deposits $= 0.10$ or 10%). It does not even matter whether the 10 percent reserve ratio is imposed by law or is merely smart profit-maximizing behaviour by the banks that balances risk and reward. The risk is the possibility of being caught short of cash; the reward is the net interest income earned.

Why were the banks able to create money? Originally, there was \$1,000 of cash in circulation. That was the money supply. When paid into bank vaults, it went out of circulation as a medium of exchange. But the public got \$1,000 of bank deposits against which cheques could be written. The money supply, cash in circulation plus bank deposits, was still \$1,000. Then the banks created deposits *not* fully backed by cash reserves. Now the public had \$10,000 of deposits against which to write cheques. The money supply rose from \$1,000 to \$10,000. *The public was willing to hold and use bank deposits as money, make payments by transferring deposits, by cheque or e-transfer, for example, and willing to borrow from the banks*. This allowed the banks to create money by making loans based on their fractional reserve ratio.

Alternatively, suppose the public loses confidence in banks and withdraws and holds more currency. The banks are still able to create deposits but the extent of the deposit creation is limited by the public's withdrawal of currency. Bank reserves are reduced. A fall in public confidence in the banks in times of financial problems and bank failures like those in that arose in the autumn of 2008 and even today in some European countries would result in a rise in the currency holdings outside banks. Bank deposits and lending capacity would be reduced as a result, and in extreme cases bank solvency might be at risk without central bank support.

Financial panics

Most people know that banks operate with fractional reserve ratios and are not concerned. But if people begin to suspect that a bank has lent too much, made high risk loans or faces problems in raising funds which would make it difficult to meet depositors' claims for cash, there would be a *run on the bank* and a **financial panic**. Recognizing the bank cannot repay all depositors immediately, you try to get your money out first while the bank can still pay. Since everyone does the same thing, they ensure that the bank is unable to pay. It holds cash equal to a small percentage of its deposit liabilities and will be unable to liquidate its loans in time to meet the demands for cash.

> **Financial panic**: a loss of confidence in banks and rush to withdraw cash.

Banking problems in Greece in the spring and early summer of 2015 provide an example. Concerns that the Greek government might default on loan payment agreements with IMF and European Union raised the possibility that Greece might leave the euro and return to its earlier national currency: the drachma. Should that happen, all euro deposits in Greek banks would convert to drachmas at an exchange rate that would reduce their real value substantially. Fearing this possibility, depositors in Greek banks tried to withdraw their balances in cash while they were still redeemable in euros.

Greek banks, like other banks, operate on a fractional reserve basis. They could not meet this 'run on the bank' without outside support and assistance. The European Central Bank provided emergency cash to the banks but the run continued. In response, limits were placed on the amount of cash a depositor could withdraw at any one time. These measures sustained the banks until a solution to Greek debt crisis was negotiated and immediate concerns about Greek membership in the euro and the value of Greek bank deposits subsided.

However, earlier experience shows how financial crises can arise in other ways. In 2008-2009 the crisis originated in the US mortgage market and real estate sector and caused wide spread problems for banks. Many banks had become reliant on large denomination, short-term deposits as sources of funds to support their mortgage lending. Other non-bank financial institutions like insurance companies and pension funds, as well as a relatively small number of individual customers bought these deposits. As the recession and falling property values emerged, the financial community began to worry that home-owners would not be able to pay back their mortgages.

If that happened banks would not be able to pay back depositors money, especially the large denomination short-term deposits. Once non-bank portfolio managers realized that it was difficult if not impossible to evaluate the risks of large denomination deposits, financial institutions that relied on renewing and issuing new deposits to raise funds were in difficulties. The supply of funds to replace expiring deposits dried up and banks could not repay depositors. Several large financial institutions in the United States and in other countries required government rescues or failed. The plight of famous names like *Bear Sterns*, *Countrywide Financial*, *Fannie May*, and *Freddie Mac* became headline news.

Banks in Canada were not immune to the financial difficulties created by the collapse of the large denomination deposit markets. All the major chartered banks were holding some. They were forced to accept that without a market these deposits would no longer be a source of funds. Fortunately, Canadian banks relied more heavily on strong smaller retail depositor bases as sources of funds. The banks remained financially strong and public confidence in the banks did not collapse. No Canadian bank failed or required a government bailout.

Fortunately, financial panics involving depositor runs on the bank are rare, particularly in Canada. A key reason for this, which we discuss in the next chapter, is that the central bank, the Bank of Canada, and other national central banks, will lend cash to banks in temporary difficulties. Furthermore, deposit insurance plans like the Canadian Deposit Insurance Corporation, CDIC, cover individual bank deposits up to $100,000 against default. Knowledge of these institutional arrangements helps prevent a self-fulfilling stampede to withdraw deposits before the bank runs out of cash.

By contrast, the financial crisis and the extended real estate and credit collapse in 2008 created large problems for US banks. Loan and financial asset defaults destroyed bank assets and bank liquidity. Even in the absence of panics and bank runs, many banks became insolvent without sufficient liquid assets to cover their liabilities. Failed bank data illustrates the scale of the problem. The US Federal Deposit Insurance Corporation lists 457 US bank failures over the period January 2008 to September 2012. In the four preceding years, January 2004 to December 2007 there were just 7 US bank failures. More stable conditions similar to those before 2008 have returned. The most recent bank failure list reports 8 bank failures in 2017, 0 in 2018, and just 1 as of May 2019.

20.5 The monetary base and the money supply

Table 20.1 showed that bank deposits are the major component of the money supply in Canada, as in most industrial countries. Bank deposits depend in turn on the cash balances held by banks and the public's willingness to hold bank deposits and borrow from the banks.

To complete our analysis of how the money supply is determined, we need to examine three things:

1. The source of the cash in the economy.
2. The amount of that cash that is deposited in the banking system, rather than held as cash balances by the public.
3. The relationship between the cash supply and the money supply that results from public and bank behaviour.

Today, in developed countries, central banks are the source of bank cash reserves. The central bank, the Bank of Canada in Canada, controls the issue of token money in the form of Bank of Canada notes. These are the $5, $10, $20, $50, and $100 bank notes you can withdraw from the bank when you wish to redeem some of your bank balance in cash. Bank reserves are mainly the banks' holdings of these central bank notes in their vaults and bank machines. Our bank deposits are now redeemable in Bank of Canada notes. The central bank has the responsibility to manage the supply of cash in the economy. We will examine the details of central bank operations in Chapter 22.

The cash the central bank provides to the economy is called the **monetary base (*MB*)** and is sometimes referred to as the stock of high-powered money. It is the legal tender into which bank deposits can be converted. It is the ultimate means of payment in transactions and the settlement of debts. Notes and coins in circulation and held by the banking system are the main part of the money issued by the central bank. As we discussed earlier, the commercial banks hold small settlement balances in the central bank to make inter-bank payments arising from cheque clearings.

> **Monetary base (*MB*)**: legal tender comprising notes and coins in circulation plus the cash held by the banks.

The public's decisions about the use of cash or banks deposits determine how much of the monetary base is held by the banks. The simple example of deposit creation in Table 20.4 assumed the public deposited all its cash with the banks. This was a useful simplification that ignores changes in the cash people hold.

Our main interest is the relationship between the money supply in the economy, *the total of cash in circulation plus bank deposits*, and the *monetary base* created by the central bank. Assuming the public holds just a small fixed amount of cash and using our earlier discussion of the fractional reserve ratio in the banking system, we can define a **deposit multiplier**. The deposit multiplier provides the link between the monetary base created by the central bank and the money supply in the economy. It also predicts the change in money supply that would result from a change in the monetary base supplied by the central bank.

> **Deposit multiplier**: the change in the bank deposits *caused* by a change in the monetary base.

$$\text{Bank deposits} = \text{deposit multiplier} \times \text{bank reserves}$$

$$\text{Deposit multiplier} = \frac{\Delta \text{ deposits}}{\Delta \text{ bank reserves}}$$

The value of the deposit multiplier depends on rr, the banks' ratio of cash reserves to total deposits. Banks' choice of a ratio of cash reserves to total deposits (rr) determines how much they can expand lending and create bank deposits based on their reserve holdings. The lower the reserve ratio (rr), the more deposits banks can create against given cash reserves, and the larger is the multiplier. This is the relationship illustrated in Table 20.4.

Similarly, the lower the non-bank public's holding of cash, the larger is the share of the monetary base held by the banks. When the banks hold more monetary base, they can create more bank deposits. The lower the non-bank public's currency ratio, the larger are bank holdings of monetary base and the larger the money supply for any given monetary base.

The money multiplier

For example, suppose banks wish to hold cash reserves R equal to a fraction rr of their deposits D. Then:

$$\text{Bank cash reserves} = \text{reserve ratio} \times \text{deposits, i.e.}$$

$$R = rrD \tag{20.2}$$

To keep the example simple assume that we can ignore the small amount of cash held by the non-bank sector. As a result, the monetary base is mainly held as cash in bank vaults and automatic banking machines. This means from Equation 20.2 that:

$$\text{Monetary base} = \text{reserve ratio} \times \text{deposits, i.e.}$$

$$MB = rrD \tag{20.3}$$

and the *deposit multiplier*, which defines the change in total deposits as a result of a change in the monetary base, is:

$$\text{Change in deposits} = \text{change in monetary base divided by the reserve ratio, i.e.}$$

$$\frac{\Delta D}{\Delta MB} = \frac{1}{rr} \tag{20.4}$$

which will be greater than 1 as long as rr is less than 1.

If, for example, banks want to hold cash reserves equal to 5 percent, and the non-bank public does not change their holdings of cash, the deposit multiplier will be:

$$\frac{\Delta D}{\Delta MB} = \frac{1}{0.05} = 20$$

When public cash holdings are constant, the deposit multiplier tells us by how much deposits (and therefore the money supply in the economy), notes, and coins in circulation (outside the banks and bank deposits) would change as a result of a change in the monetary base. In this example, a \$1 change in the monetary base results in a change in deposits and the money supply equal to \$20.

$$\text{Money supply } (M) = \text{cash in circulation} + \text{bank deposits } (D)$$

$$M = \text{cash} + R/rr \tag{20.5}$$

We can see from the way we have found the deposit multiplier that it depends on the decisions made by the banks in terms of their reserve holdings. For simplicity the public holds a fixed amount of cash in addition to bank deposits as money. If you experiment with different values for *rr*, you will see how the deposit and money multiplier would change if the reserve ratio were to change. Furthermore, if the public were to change their cash holdings the cash reserves available to the banks would change and the deposit multiplier would cause a larger change in the money supply.

The importance of bank reserve decisions and public cash holdings decisions is illustrated by recent financial conditions in Europe. As a result of banking crisis and bailouts during and after the financial crisis of 2008, the public had concerns about the safety of bank deposits and decided to hold more cash. At the same time banks found it difficult to evaluate the credit worthiness of potential borrowers and the risks involved in short-term corporate lending or junior government bonds. The public's cash holdings increased and the banks increased their reserve ratios. These shifts in behaviour would reduce the money supply, making credit conditions tighter, unless the central bank provided offsetting increases in the monetary base.

How big is the deposit multiplier?

Now that we have a formula for the deposit multiplier, we can ask: What is the size of the multiplier in Canada? Based on data in Table 20.1 above, in September 2018, the monetary base was \$94.2 billion, and the money supply defined as M1B was \$878 billion. These data suggest a bank reserve ratio with respect of M1B $= \$94.2/\$878$ which is approximately 10.7 percent giving a deposit multiplier of $1/0.107$.

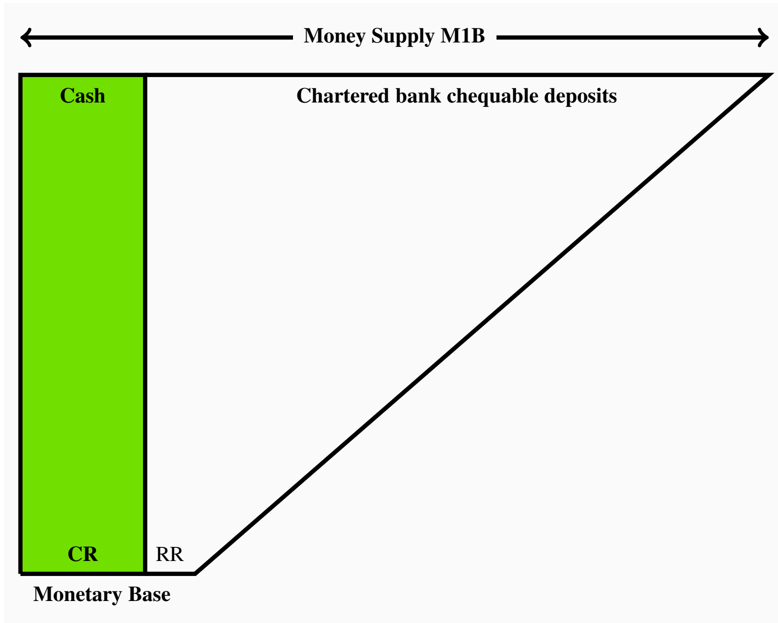
$$\frac{\Delta D}{\Delta MB} = \frac{\$878}{\$94.2} = 9.3$$

Each \$100 change in monetary base would change the money supply by about \$930.

However, using a broader definition of money supply such as 'currency outside banks and all chartered bank deposits' gives a Canadian money supply of \$1,632.6 and a deposit multiplier of \$1,632.6/\$94.2 = 17.33.

Figure 20.1 summarizes the relationship between the monetary base and the money supply. It shows the monetary base used either as cash in circulation or held as cash reserves by the banks. Since banks operate with fractional reserve ratios, banks have the leverage to expand the money supply through their lending and deposits creation based on their reserves *RR*. We also see that the money supply is heavily dependent on the size of the monetary base, reserve ratios used by the banks and the willingness of the public to hold bank deposits and the willingness of the banks to lend.

Figure 20.1: The monetary base and the money supply



The explanation of banking and the money supply in this chapter provides the money supply function we will use in the next chapter. It is combined there with a demand for money function in the money market to determine the equilibrium rate of interest. That rate of interest integrates money and financial markets with the markets for goods and services in aggregate demand.

A simple money supply function illustrates the determinants of the money supply. The three key variables are:

1. MB, the monetary base;
2. the public's holdings of cash; and
3. rr, the banks' reserve ratio.

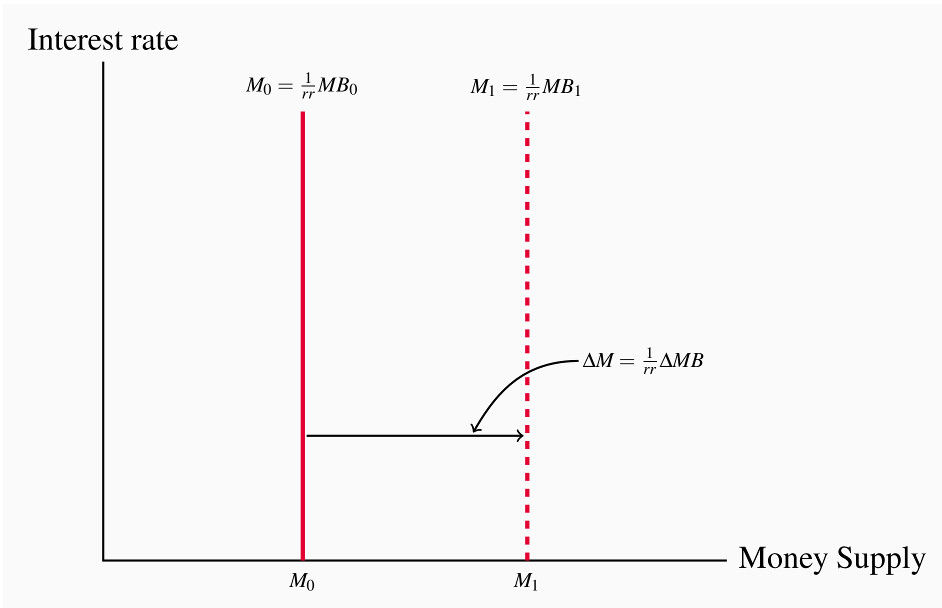
Using Equation 20.5 above, where M is the money supply, we can write:

$$M = \frac{1}{rr} \times MB \tag{20.6}$$

The central bank's control of the monetary base, MB, gives it control of the money supply, M, as long as cash holdings and rr are constant.

Figure 20.2 uses a diagram to illustrate the money supply function and changes in the money supply. The line M_0 shows the size of the money supply for a given monetary base MB_0 and the money multiplier. The money supply in this diagram is vertical, because we assume cash holdings and the reserve ratio are not affected by the interest rate. M is therefore independent of the nominal interest rate i, which is measured on the vertical axis. This is the supply side of the money market with quantity measured on the horizontal axis and interest rate, which is analogous to price, on the vertical axis.

Figure 20.2: The money supply function



The vertical M_1 illustrates the increase in the money supply as a result of an increase in monetary base from MB_0 to MB_1, working through the money multiplier. A reduction in the monetary base would shift the M line to the left based on the same relationship.

Monetary policy

The central bank conducts monetary policy through its control of the monetary base. The money supply function shows us how, *if rr is constant, the central bank's control of the monetary base gives it the power to change money supply and other financial conditions in the economy*. If the central bank increases the monetary base, banks have larger cash reserves and increase their lending, offering favourable borrowing rates to attract new loans and create more deposits. In Figure 20.2 the increase in the monetary base to MB_1 causes an increase in money supply (ΔM) by the change in MB (ΔMB), multiplied by the money multiplier. The money supply function shifts to the right to M_1. A decrease in the monetary base would shift the M function to the left, indicating a fall in the money supply.

Next

Now that we have examined money, the banking system, and the size of the money supply, we have one important side of the *financial market* that will link money to expenditure and economic activity. This is the supply side of the market. In the next chapter, we will study the reasons why people wish to hold money balances. We also study how the portfolio choices people make between money and other assets create the demand for money balances. The interaction between the supply of money balances and the demand for money balances determines the prices of financial assets and interest rates. Interest rates in turn provide an important link between money, financial markets, and expenditures in markets for goods and services, both directly and through the foreign exchange rate.

Key Concepts

Money has four functions: a **medium of exchange** or means of payment, a **store of value**, a **unit of account**, and a **standard of deferred payment**. The medium of exchange function distinguishes money from other assets.

In a **barter economy**, trading is costly because there must be a double coincidence of wants. Money, a medium of exchange, reduces the costs of exchange and allows resources to be used for other things.

A **token money** is a convertible claim on commodity money. Because its monetary value greatly exceeds its production costs, token money economizes on the resource costs of transactions.

Fiat money is money the government has declared **legal tender**. The **central bank** controls the supply of legal tender.

Fintech is application of computer and internet technologies to facilitate the flow of payments and financial information.

E-payments use debit cards, telephone banking, internet banking, e-transfers, and mobile banking, which provide access to manage deposit accounts and make payments.

E-monies are multi-purpose pre-paid payment cards, and stored-value cards and gift cards for purchases at specific companies.

Cryptocurrencies are a form of decentralized e-money. They are financial assets but have no centralized issuer and are not denominated in any national currency. They are not money in that they do not serve as a *'generally acceptable'* means of payment. *Bitcoin* and *Ethereum* are examples.

The Canadian **money supply** is the sum of currency in circulation outside the banks and bank deposits.

The **monetary base** is comprised of notes and coins in circulation plus cash held by banks.

The Canadian banking system is made up of a **central bank** and a number of **commercial banks** and other institutions called **near banks**.

Banks are **financial intermediaries**. Bank deposits, which can be transferred by cheque or debit card, provide a convenient means of payment. Bank services plus interest payments on deposits attract funds into the bank. Banks use these funds to make loans, purchase securities, and finance expenditures. The general acceptance of bank deposits as money, and well-developed financial markets, allow modern banks to operate with very low cash reserve ratios.

Clearing balances: cash balances used by a bank to settle the differences between its customers' payments and receipts involving other banks.

Banks create money by making loans and creating deposits based on a **fractional cash reserve ratio,** rr. The banks' reserve ratio involves a trade-off between earnings and **bankers' risk**.

The **monetary base** MB is currency in circulation plus banks' cash reserves. The **money multiplier** is a ratio of a change in the money supply to the change in the monetary base that caused it, $\Delta M/\Delta MB$. The money multiplier is larger the smaller is the cash reserve ratio of the banks, rr.

The **money supply**, M, is currency in circulation plus bank deposits. The size of the money supply is determined by the monetary base, MB, the banks' cash reserve ratio, rr, and the private sector's cash holdings, when cash holdings are constant. From Equation 20.5:

$$M = \text{cash} + \frac{1}{rr} \times MB$$

Exercises for Chapter 20

Exercise 20.1 What are the functions of money? What is money in Canada today? What is the money supply in Canada today? Are debit cards and credit cards money?

Exercise 20.2 Since both central banks and commercial banks can create money what is the key difference between a central bank, like the Bank of Canada, and the many commercial banks in the financial industry?

Exercise 20.3 Suppose the banks receive \$100 cash from a new deposit of funds previously held outside the banking system. If banks operate with a 5% cash reserve ratio, use simple balance sheets to show by how much this new cash would affect lending and deposits of all banks in the system.

Exercise 20.4 If banks have a 10% cash reserve ratio how much lending and deposit creation can they undertake after they receive a new \$1,000 cash deposit? Would it be in the banks' interest to find ways to reduce any cash balances the public holds? Why?

Exercise 20.5 What protection does the Canadian Deposit Insurance Corporation provide for your money if your bank is unable to pay cash to its depositors?

Exercise 20.6 Define the deposit multiplier and explain how it might be used.

Exercise 20.7 Suppose a crisis in financial markets, like the collapse of public willingness to hold large denomination term deposits in 2007 and 2008, increases the risk banks attach to lending and the non-bank public attaches to all bank deposits. What are the implications for the desired cash reserve ratio, the money supply multiplier, and the money supply?

Exercise 20.8 Using a diagram illustrate and explain the determinants of the position and slope of the money supply function assuming an initial monetary base of \$1,000 when $rr = 5\%$. If the monetary base were to increase by 10% how would the money supply and the money supply function in your diagram change?

Chapter 21

Financial markets, interest rates, foreign exchange rates & AD

In this chapter we will explore:

Interest rates are everywhere. They provide income to lenders and impose costs on borrowers. Many types of bank and near-bank deposits pay interest income to their holders, as do guaranteed investment certificates (GICs) and bonds issued by businesses and governments and other income paying assets. These are assets to holders and liabilities to issuers in terms of interest payments and redemptions.

Households, businesses and governments borrow using consumer loans, credit cards, lines of credit, mortgages, leases, bonds and other forms of debt. The interest rate attached to these borrowings is a cost of financing expenditures not covered by current income.

The interest rate links financial markets to markets for goods and services. But what is 'the interest rate'? There are different interest rates attached to different forms of borrowing. There are different spectrums of interest rates based on the terms of the loan and the perceived risk of default. However, it is usually the case that interest rates across a spectrum rise or fall together. Changes in 'the interest rate' describe these shifts up or down across the spectrum for all assets.

This chapter addresses two key questions:

1. How is the interest rate determined?
2. How does the interest rate affect aggregate demand?

Explaining why the public decides to hold some wealth in money balances rather than other types of financial assets is the first step.

21.1 Portfolio choices between money and other assets

A financial portfolio is a collection of financial assets. It might include money balances, bonds, equities, mortgages, and mutual funds. The structure of a portfolio, the proportion held in each type of asset, reflects two main characteristics of the assets involved:

1. The *returns* paid by different financial assets
2. The *risks* arising from changes in the market prices of assets

Wealth holders and institutional portfolio managers for pension funds and insurance companies like their portfolios to pay high returns with low risk. To achieve this, they hold mixed portfolios of money and other financial assets.

Suppose you win $10 million in a lottery. Now that you have wealth, what are you going to do with it? You will no doubt spend some and give some away. That is a wealth effect, but what about the balance of your winnings? You have to make a portfolio choice. Will you hold your wealth as money in the bank? Will you put your money in the stock market? Will you put your money in the bond market? Will you buy bitcoin?

If you consult a financial planner, he or she will probably recommend a mixed portfolio made up of money, bonds, and equities. That recommendation will be based on your intention to increase your wealth and draw income from it while protecting it from losses in financial markets.

Money holdings are an important part of the portfolio. Money is the medium of exchange. It can be used directly to make payments for goods and services or to settle debts. Other assets, for example bonds, cannot be used as a means of payment. Furthermore, money has a fixed nominal price. It is a "safe asset." Wealth held as money does not rise or fall with the rise or fall in financial asset prices on stock and bond markets. However, money is exposed to the risk that inflation will lower its real purchasing power.

Other financial assets differ from money in three respects. First, they cannot be used as a means of payment. To use them to make a payment you would first have to sell them for money, at their current market price, and then use the money to make the payment. Second, they offer a return in the form of an interest payment, a dividend payment, or a rise in price that provides income to the portfolio holder. Third, because the prices of financial assets like bonds or stocks fluctuate daily on financial markets, these assets carry the risk that their values may decline significantly from time to time.

Portfolio management recognizes these differences between assets by trading some return for lower risk and greater convenience in the mix of assets held. Money in the portfolio offers the convenience of the means of payment, providing low risk but zero return. Other assets offer a flow of interest and dividend income, and possible capital gains if asset prices rise, but the risk of capital loss if prices fall.

Movements in the S&P TSX, an index of the value of approximately 250 companies, in the first quarter of 2020 illustrates the risks in holding equities rather than cash. The index stood at 17,099 at the beginning of January 2020. Then, as a result of the economic impact of COVID-19, it dropped to 11,228 by late March 2020, a capital loss of about 34 percent. It then recovered to 14,967 on May 8, 2020, a capital gain of 33 percent from the low in March but still 12.5 percent below its January value. If you had held a portfolio of cash or bonds instead of equities you would have avoided all this excitement.

This portfolio choice between money balances and other assets is the basis for our discussion of the demand for money balances in the remainder of this chapter.

Bond prices, yields and interest rates

The demand for money comes from an understanding of the relationship between interest rates, bond coupons, the prices of financial assets, and yields on financial assets. To keep the examples simple assume only one type of financial asset, a bond. However, the prices and yields of other financial assets are related to interest rates in the same way as bond prices.

Several basic concepts and definitions are important. A **bond** is an asset that makes one or more *fixed money payments* to its holder each year until its maturity date. On its maturity date, it also repays its principal value. Governments and businesses issue and sell bonds on financial markets to raise funds to finance expenditures.

> **Bond**: a financial contract that makes one or more fixed money payments at specific dates in the future.

The **interest rate** is the *current* market rate, expressed as a percentage, paid to lenders or charged to borrowers.

> **Interest rate**: the current market rate paid to lenders or charged to borrowers.

A **bond coupon** is the *fixed money payment* made *annually* to the holders of the bond from the date of issue until the date of maturity. The coupon rate is a fixed percentage of the principal value of the bond at the time of issue. For example a 3% bond pays $3.00 annually per $100 of principal until its maturity date.

> **Bond coupon**: the *annual* fixed money payment paid to a bond holder.

The **price of a marketable bond** is the current price at which it can be bought or sold on the open bond market at any time between its date of issue and its maturity date.

> **Price of a marketable bond**: the current price at which the bond trades in the bond market.

The **yield on a bond** is the return to a bond holder expressed as an annual percentage rate, which is a combination of the coupon payments and any change in the market price of the bond during the period in which it is held.

> **Yield on a bond**: the return to a bond holder expressed as an annual percentage.

Bond prices depend on current market interest rates. The **current price of a bond** is the present value of the future payments it will provide. The **present value** is the *discounted* value of those future payments. It recognizes that money payments in the future are worth less than money payments today.

> **Bond price**: the *present value* of future payments of interest and principal.
>
> **Present value** is the *discounted* value of future payments.

To help understand present value, ask the following question: If someone promises to give you \$1,000, would you rather have it today or a year from today? Notice that \$1,000 lent at an interest rate of 3% (i.e. $3/100 = 0.03$) would give you a sum of:

$$\$1,000 \times (1 + \text{the interest rate}) = \$1,000 \times (1.03) = \$1,030$$

one year from today. In the same way, the amount of money you need to lend today to have \$1,000 one year from today is:

$$\begin{aligned} \$M \times (1.03) &= \$1,000 \\ \$M &= \$1,000/1.03 \\ \$M &= \$970.87 \end{aligned}$$

When the market rate of interest is 3 percent, the present value of \$1,000 to be received one year in the future is \$970.87.

Experimenting with different interest rate assumptions in this present value calculation illustrates that the present value of \$1,000 to be paid one year from today changes with the rate of interest. Higher interest rates reduce present values while lower rates increase them. For example, if the current market rate is 5% the present value of \$1,000 to be received one year from today is:

$$\begin{aligned} \$M \times (1.05) &= \$1,000 \\ \$M &= \$1,000/1.05 \\ \$M &= \$952.38 \end{aligned}$$

This relationship is the key to understanding bond prices and how they fluctuate over time. A rise in market interest rates lowers the present value of fixed future payments. A fall in market rates increases present values of fixed future payments.

In general, because the future payments offered by bonds are fixed in dollar terms, the prices of marketable bonds vary inversely to market rates of interest. Rising interest rates mean falling bond prices, and falling interest rates mean rising bond prices. There are many types of bonds that differ by coupon, maturity date, frequency of future payments, and in other ways. However, the relationship between prices, yields, and interest rates remains the same. Because bond prices are the present value of future payments, prices and interest rates move in opposite directions.

Furthermore, the size of the change in the price of a bond as a result of a change in the interest rate depends on the bond's term to maturity. The prices of longer-term bonds are more volatile than those of shorter-term bonds. This an important consideration for bond portfolio managers concerned with trade-offs between risk and return.

Asset markets like the bond market are very active. Large volumes of bonds are bought and sold every business day. Example Box 21.1 at the end of the chapter gives a more detailed example of the relationship between market interest rates, bond prices and yields to maturity. You can see the recent movements in yields on Canadian government bonds issued with different terms to maturity and coupon yields at: `https://www.rbcdirectinvesting.com/pricing/gic-bond-rates.html`

21.2 The demand for money balances

Canadians held M2 money balances of $1,716 billion in May 2019. Three variables that may explain the size of these holdings are: the interest rate, the price level, and real income. Together they provide the basis for a theory of the demand for money.

Why hold money?

It is important to distinguish between money and income when discussing the demand for money. You might have a high income but no money, or no income and lots of money. That is because income is a flow of funds over a period of time. If you spend your income as it is received you will not accumulate a *stock of money*. Alternatively, you might have a stock of money or a money balance but no income. Then you can choose to either hold or spend your money. If you have no income you can finance a flow of expenditures by spending your money balance.

In Chapter 20, money was a means of payment and a store of value. Those two functions motivate the demand to hold at least some wealth in money balances. There are alternative stores of value. Bonds, equities, precious metals, real estate, and art are a few examples. The quantity of money people choose to hold is part of the *portfolio decision* they make about their wealth. They choose money instead of some other asset.

To develop the demand for money balances it is useful to simplify the portfolio decisions by assuming there are only two assets:

1. *Money*, which has a constant money price, pays no interest income but does serve as the means of payment.
2. *Bonds*, representing all interest-earnings assets, have money prices that change if market interest rates change, but are not means of payment.

The financial wealth people build up by saving some of their income calls for a decision. People could hold this wealth as money, which pays no interest, but is a *safe asset* because its price is constant. Or they could hold this wealth in bonds, which pay interest income but are *risky* because bond prices move up and down as market interest rates move down and up. If the *expected return* to holding bonds is positive (due to the interest rate together with any change in price) why would people hold any money balances?

The demand for money comes in three parts, namely:

1. The transactions demand;
2. The precautionary demand; and
3. The asset or speculative demand.

The transactions demand

As the name suggests, the transactions demand for money is based on money being the means of payment. People and businesses hold some money to pay for their purchases of goods, services and assets. This demand reflects the lack of coordination of receipts and payments. Income is paid bi-weekly or monthly but purchases are made more frequently and in smaller amounts. Pocket money and bank balances that can be transferred by debit card are readily available to make these purchases between paydays. If all income receipts were used on paydays to buy bonds to earn interest income it would be costly and inconvenient to sell bond holdings bit by bit as payments were made. The costs of frequent switching between money, bonds and money would more than offset any interest income earned from very-short-term bond holdings.

The precautionary demand

Uncertainty about the timing of receipts and payments creates a precautionary demand for money balances. There are two sides to this uncertainty. On one side there may be some unexpected changes in the timing or size of income receipts. Regular payments can still be made if enough money is available, over and above that need for usual expenses and payments. Alternatively, unexpected or emergency expenses in terms of appliance, computer or car breakdowns or unexpected opportunities for bargains or travel can by covered by precautionary money holdings. Money balances cover the unexpected gaps between income receipts and payment requirements without the costs and inconvenience of selling bonds on short notice.

The asset or speculative demand

The asset or speculative demand comes from financial portfolio decisions rather than the lack of coordination and uncertainty behind the two preceding demands. Businesses and professional portfolio managers use money balances to take advantage of expected changes in interest rates. Essentially they speculate by switching between bonds and money based on their own forecasts of future interest rates.

Recall that bond prices and interest rates vary inversely. If while holding money balances you predict a fall in interest rates, you buy bonds. If your prediction is right and interest rates do fall, the prices of your bonds rise. Now you can sell and harvest the capital gain you earned by speculating in the bond market. Alternatively, if you correctly predict a rise in interest rates and act before it happens you can avoid a capital loss on your bond holdings by selling and holding money before the interest rate rises.

Even if portfolio managers are not interested in speculating on interest rate changes there is an asset demand for money. A mixed portfolio of money and bonds is less risky than one that holds only bonds. The money component has a stable market price while the bond component provides interest income along with the risk of a variable price. Changing the shares of money and bonds in the portfolio allows the manager a trade-off between return and risk. However, as interest rates rise, the opportunity cost of holding a share of the portfolio in money rises. Furthermore, the estimated risk from the bond share of the portfolio may fall if interest rates are expected to fall in the near future. As a result, rising interest rates reduce the asset demand for money balances.

The demand for money function

The demand for money balances is summarized by a simple equation. Let the size of the real money balances people wish to hold for transactions and precautionary reasons (L_t) be a fraction k of GDP. With nominal GDP defined as real GDP (Y) times the GDP deflator (P), nominal GDP $= PY$. Using this notation, the demand for nominal money balances for transactions and precautionary reasons is kPY, and the demand for real balances is kY, where k is a positive fraction. When real income changes, bringing with it changes in spending, the change in the demand for real money balances changes is determined by k. This makes a link between part of the demand for money balances (L_t) and income, namely $L_t = kY$.

What is the value of k in Canada? In the second quarter of 2019, Canadians held money balances as measured by M2 of about \$1,716 billion. Nominal GDP in that quarter was \$2,250 billion measured at an annual rate. If we divide M2 holdings by GDP, we get $k = \$1,716/2,250 = 0.76$, or about 76 percent of annual income. This value of k suggests that a rise in GDP of \$100 will increase the demand for money balances by \$76, measured in either nominal or real terms.

Changes in nominal interest rates also change the size of the money balances people wish to hold, based on the asset motive. A rise in interest rates increases the opportunity cost of holding money balances rather than bonds. It may also create the expectation that interest rates in the future will fall back to previous levels. As a result, people will want to use some of their money balances to buy bonds, changing the mix of money and bonds in their wealth holdings. A fall in interest rates has the opposite effect.

The way people adjust their portfolios in response to changes in interest rates results in a negative relationship between the asset demand for money balances and the nominal interest rate. Then using $-h$ to measure the change in money balances in response to a change in interest rates can be written as $-h = \Delta L/\Delta i$. If individual and institutional portfolio mangers' decisions are very sensitive to the current interest rates, h will be a large negative number. A small rise in interest rates will cause a large shift from money to bonds. Alternatively, if portfolio decisions are not at all sensitive to interest rate changes, h would be zero.

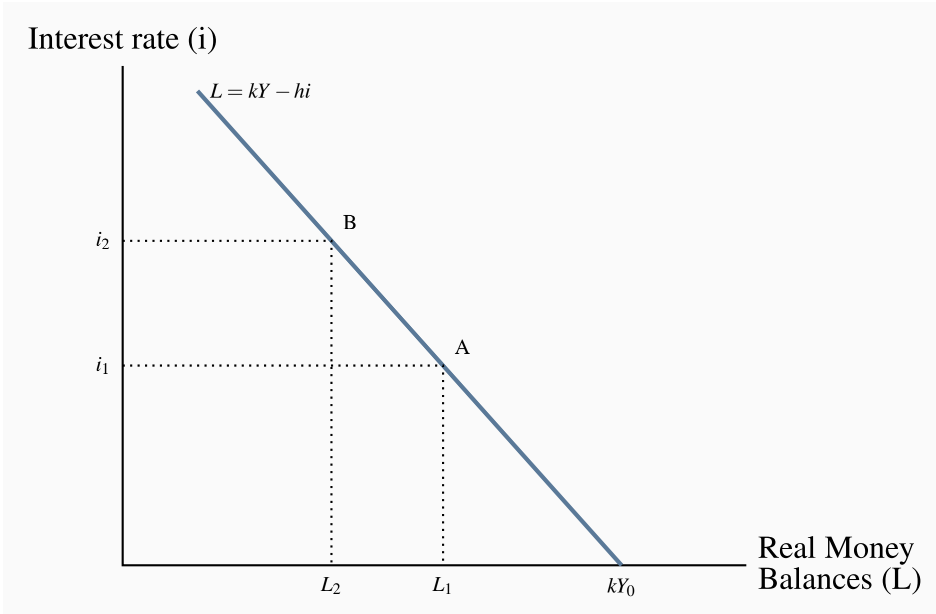
Putting these components of the demand for real money balances together gives the demand for money function, which is a demand to hold real money balances L:

$$\begin{aligned}\text{Money balances demand} \;=\;& \text{transactions and precautionary balance based on income } (kY) \\ &+ \text{the asset demand based on interest rates } (-hi)\text{, i.e.:}\end{aligned}$$

$$L = kY - hi \tag{21.1}$$

Figure 21.1 shows the relationship between the demand for real money balances and the interest rate, drawn for a given level of real GDP, Y_0. The demand for money function would have an intercept of kY_0 on the horizontal axis. At higher interest rates the opportunity cost of holding money balances is higher because the expected return from holding bonds is positive. The negative slope of the demand function shows how people change their demand for money when interest rates change. The slope of the demand curve for money is $-1/h$. The effect of a change in the interest rate is shown by a movement along the L function. A change in real income would require us to draw a new demand for money function, to the right of L_0 if Y increased, or to the left if Y decreased.

Figure 21.1: The demand for real money balances



This straight line demand for money function is a useful simplification. However, it would be more realistic to draw the function with a decreasing slope as interest rates decline. That would capture two important ideas.

First, as interest rates fall, and fall relative to the costs of buying and selling bonds, opportunity costs decline faster than interest rates.

Second, consider the speculative demand for money. As interest rates fall, the riskiness of bonds increases. A subsequent rise in interest rates has a larger negative effect on bond prices. Expectations of future increases in interest rates may strengthen as interest rates decline. As a result, portfolio managers may shift funds increasingly from bonds to money as interests fall. If their expectations are confirmed by events they avoid the capital losses caused by falling bond prices.

21.3 Financial market equilibrium & interest rates

The money supply and the demand for money in the financial market determine nominal interest rates. From Chapter 20 a nominal money supply depends primarily on the monetary base and the money multiplier, namely:

$$M = \frac{1}{rr} \times MB(20.6)$$

The demand for money is a demand for *real* money balances as determined by real income and interest rates.

$$L = kY - hi$$

The **real money supply** is simply the nominal money supply M divided by the price level P, (M/P), which measures its purchasing power in terms of goods and services.

> **Real money supply** (M/P): the nominal money supply M divided by the price level P.

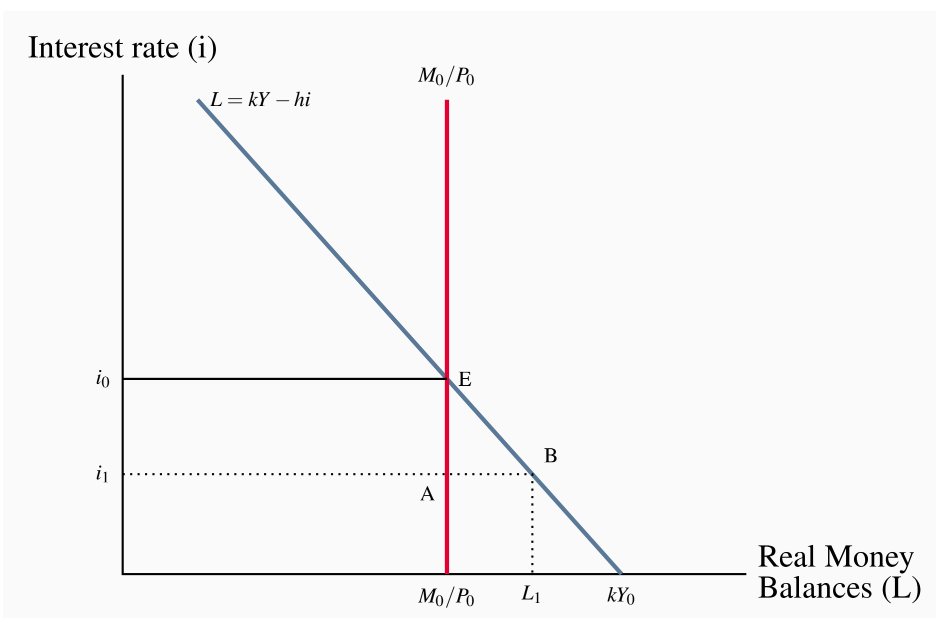
The central bank, as the source of the monetary base MB, can control the nominal money supply, *as long as the banks' cash reserve ratio rr and the public's holdings of cash are constant*. The next chapter explains how the central bank manages the monetary base. If the price level is fixed, the central bank also controls the real money supply. Changes in nominal money tend to lead eventually to changes in prices. However, the central bank can still control the real money supply in the short run—it can change M faster than prices P respond—but in the long run other forces determine real money M/P. For the moment, assume the price level as fixed.

The demand for money as a demand for real money balances is summarized above. The quantity of real money demanded rises when real income rises, but falls when nominal interest rates rise.

Money market equilibrium

Figure 21.2 combines the demand curve for real money balances from Figure 21.1 with the money supply function in Figure 20.2 to give a money market diagram. The demand curve is drawn for a given level of real income, Y_0, and the supply curve for a given monetary base MB_0. With a given price level, the central bank controls the supply of nominal and real money. The supply curve is vertical at M_0/P_0. Equilibrium in the money market is at E. At the interest rate i_0, the real money balances people wish to hold just equal the money supplied by the central bank and the banking system.

Figure 21.2: Equilibrium in the money market



To see how this market operates, suppose the interest rate is i_1, lower than the equilibrium level i_0. There is excess demand for money in the amount AB in the diagram. People want to hold money balances equal to B at the interest rate i_0, but only A is available. How does the market adjust to remove this excess demand? The answer lies in the portfolio decisions that distribute wealth between money holdings and bonds.

Consider the interaction between the bond and money markets. When portfolio managers want to restructure their holdings of bonds and money they do so by buying or selling bonds on the bond market. Their actions cannot change the supply of money balances. That is fixed by the monetary base and the money supply multiplier. As a result, bond prices and interest rates change to maintain money market equilibrium.

In Figure 21.2 the excess demand for money at the interest rate i_1 will result in a rise in interest rates. With an excess demand for money, people sell bonds to adjust their money balances. There is an excess supply of bonds. Bond prices fall. Lower bond prices mean higher bond yields and interest rates, as you will recall from our earlier discussion of asset prices and yields. The higher interest rates reduce both the excess supply of bonds and the excess demand for money. The money market adjusts by *moving along* the L curve from B to E, as people want smaller money balances relative to their bond holdings at higher interest rates.

This inverse interest rate – bond price relationship is the key to adjustments in the money market caused by changes in either the demand for or supply of money balances. Those adjustments involve trades in bonds that change bond prices and interest rates to maintain money market equilibrium.

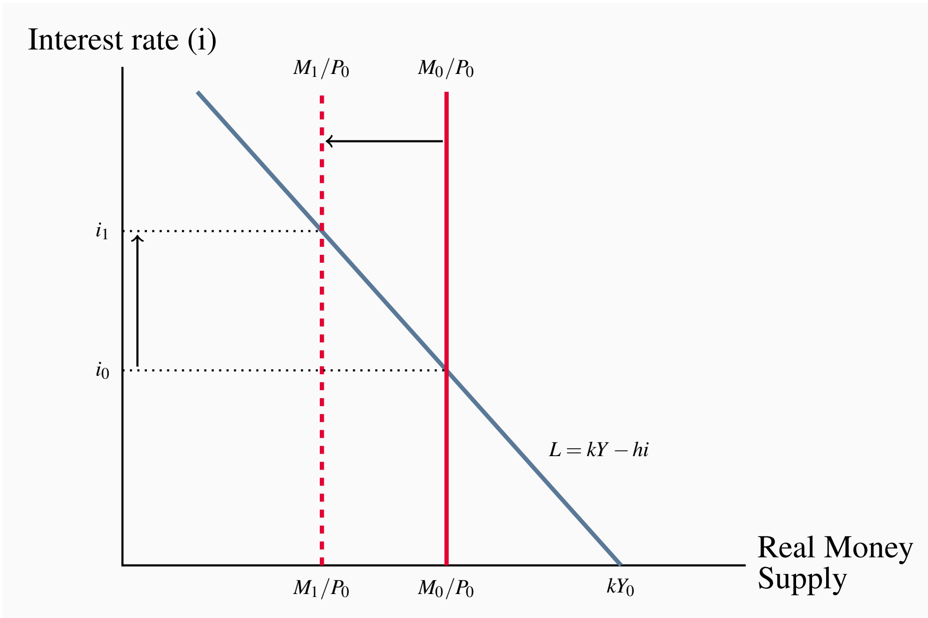
Changes in financial market equilibrium

A shift in either the money supply or money demand changes equilibrium in the money market (and the bond market). Interest rates move to restore equilibrium. Figures 21.3 and 21.4 give examples.

The effect of a change in the money supply

Suppose the central bank lowers the monetary base and the money supply contracts. For a fixed price level, lower nominal money reduces the real money supply. Figure 21.3 shows this leftward shift in the money supply curve from M_0/P_0 to M_1/P_0. The equilibrium interest rate rises from i_0 to i_1 as people sell bonds. A higher interest rate reduces the quantity of real money balances demanded, moving along the demand curve $L(Y_0)$, bringing quantity of balances demanded into line with the reduced supply. Hence, a lower money supply raises equilibrium interest rates. Conversely, a rise in the money supply lowers the equilibrium interest rate.

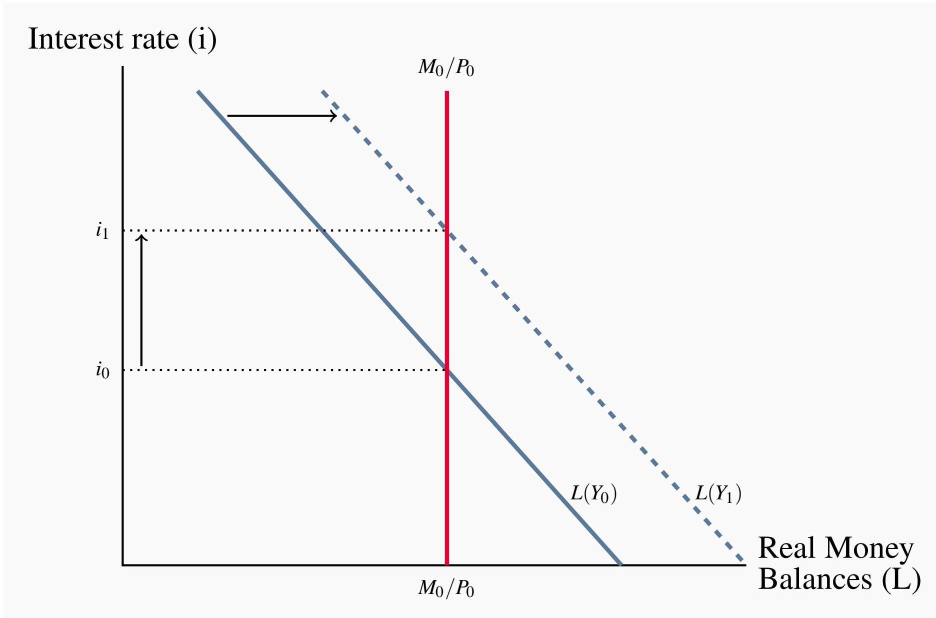
Figure 21.3: Effect of a fall in the money supply



The effect of a change in real income

Figure 21.4 shows real money demand $L(Y_0)$ for the real income Y_0. A rise in real income increases the quantity of real money balances demanded at each interest rate, shifting the demand for money function from $L(Y_0)$ to $L(Y_1)$. The equilibrium interest rate rises as portfolio managers sell bonds in an attempt to increase their money holdings. The rise in the interest rate lowers the quantity of real balances demanded, moving along the money demand function $L(Y_1)$, and keeps demand for money equal to the unchanged supply. Conversely, a fall in real income would shift the demand for money to the left and reduce the equilibrium interest rate.

Figure 21.4: Effect of a rise in real income



21.4 Foreign exchange rates

The interest rates determined in the money market have important effects on the **foreign exchange rate**. With free international trade in financial assets, portfolio managers, having chosen to hold some part of their portfolios in bonds, have an additional choice. They can hold some bonds issued by domestic borrowers and some issued by foreign borrowers. They might, for example, hold some bonds issued by the Government of Canada, some issued by the United States Treasury and some issued by other governments. Similarly, residents of other countries can choose to include bonds issued by the Government of Canada in their holdings. These choices are made on the basis of the yields on bonds established by conditions in different national money and bond markets.

Foreign exchange rate: the domestic currency price of a unit of foreign currency.

To achieve the highest return on the bond portion of their portfolios, managers buy bonds that offer the highest rate of return for a given level of risk. If interest rates are constant in other financial markets, a rise in Canadian interest rates and bond yields makes Canadian bonds more attractive to both domestic and foreign bondholders. The demand for Canadian bonds increases. A fall in Canadian interest rates has the opposite effect.

Bonds are issued and priced in national currency. Most Government of Canada bonds are denominated in Canadian dollars. US Treasury bonds are denominated in US dollars, and bonds issued by European governments are denominated in euros. If Canadians want to purchase bonds on foreign bond markets they need foreign currency to make payment. Similarly, if residents of other countries want to purchase Canadian bonds they need Canadian dollars to make payment. These *foreign exchange* requirements for trading in financial assets are the same as those for trading in goods and services. The foreign exchange market is the market in which currencies of different countries are bought and sold and foreign exchange rates are determined.

Foreign exchange markets and rates are examined in detail in Chapter 24. For now it will be enough to consider the effects of changes in domestic interest rates on the foreign exchange rate.

Consider an *increase* in the domestic money supply in Canada. Money and bond market adjustments to this increased money supply lower the Canadian interest rate. At these lower interest rates domestic bond yields are lower relative to foreign bond yields than they were before. This provides the incentive for domestic portfolio managers to switch their purchases from domestic bonds to foreign (US) bonds. To pay for foreign bonds they need foreign currency. The demand for US dollars increases.

Simultaneously, bond holders in the US shift their purchases from the now relatively low-yield Canadian bonds to US bonds. Lower sales of Canadian securities in the US market reduce the supply of US dollars. This drop in the supply of US dollars combined with the increase in demand raises the Canadian dollar price of U.S dollars. This is a **depreciation of the Canadian currency**. If the initial exchange rate was \$1.20Cdn = 1.00US, the exchange rate would be somewhat higher, say \$1.25Cdn = \$1.00US

> **Depreciation of the national currency**: a decline in the value of the currency relative to other national currencies, which results in a rise in the domestic price of foreign currencies.

In this example a fall in domestic interest rates, other things constant, causes depreciation in the domestic currency relative to foreign currencies. This interest rate-exchange rate linkage is symmetrical. Rises in domestic interest rates cause **appreciation of the national currency.**

> **Appreciation of the national currency**: an increase in the value of the currency relative to other national currencies, which results in a fall in the domestic currency price of foreign currencies.

A *decrease* in the money supply or a change in the demand for money with a fixed money supply would affect the foreign exchange rate through the same linkages. Changes in domestic financial markets and foreign exchange markets happen simultaneously. With current communications and information technology these markets adjust very rapidly and continuously. The changes in interest rates and foreign exchange rates that result from changes in domestic money market conditions have important effects on aggregate expenditure and aggregate demand.

21.5 Interest rates, exchange rates, and aggregate demand

Interest rates and exchange rates link the changes in money and financial markets to the expenditure decisions that determine aggregate demand.

The impact of financial markets, interest rates, and exchange rates on aggregate expenditure, aggregate demand, and real output, is described by the **transmission mechanism**. It has three important channels, namely:

1. the effect of interest rate changes on consumption expenditure;
2. the effect of interest rate changes on investment expenditure; and
3. the effect of interest rate changes on foreign exchange rates and net exports.

> **Transmission mechanism**: links money, interest rates, and exchange rates through financial markets to output and employment and prices.

Interest rates and consumption expenditure

The basic consumption function in Chapter 18 was illustrated by a straight line relating aggregate consumption to disposable income. The positive slope of that line, the marginal propensity to consume, showed the change in consumption expenditure that would result from a change in disposable income. The vertical intercept of the consumption function showed *autonomous consumption* expenditure, the consumption expenditure *not* determined by disposable income. Changes in income moved households *along* the consumption function. Changes in autonomous consumption expenditure changed the vertical intercept, *shifting* the consumption function *up or down*.

Changes in interest rates affect autonomous consumption expenditure in two ways.

1. Through a **wealth effect** from changes in the prices of financial assets; and
2. Through a **cost of credit** effect.

> **Wealth effect**: the change in expenditure caused by a change in real wealth.
>
> **Cost of credit**: the cost of financing expenditures by borrowing at market interest rates.

The market prices of bonds are the present values of expected future interest and principal payments. Current market interest rates are the key factor in this relationship. Similarly, interest rates and the expected stream of profits and dividend payments determine the present values and prices of company shares. Lower discount rates give higher present values and higher interest rates reduce present values. As a result, falling interest rates raise financial asset prices and rising interest rates reduce financial asset prices.

This means that changes in interest rates, by changing prices of financial assets, change the wealth held in household portfolios. A fall in market interest rates raises household financial wealth which increases household consumption expenditure. Autonomous consumption expenditure increases as a result of this **wealth effect**. A rise in interest rates would reduce autonomous consumption expenditure.

Changes in interest rates, for example a fall in rates, also has important effects on house prices by lowering the **cost of credit**, increasing the present values of rental incomes, and increasing the market prices of residential real estate. Households often use the increased market values and equity in their housing to set up home equity lines of credit with relatively low borrowing rates. This borrowing is used to finance other expenditures. Autonomous consumption expenditures change as interest rate changes change the cost and extent of this financing.

Thus, two forces—wealth effects, and availability and cost of credit—explain the effects of money on planned consumption expenditure. This is one part of the *transmission mechanism* through which money and interest rates affect expenditure. Operating through wealth effects and the supply and cost of credit, changes in money supply and interest rates *shift* the consumption function. We can recognize the effects of both income and interest rates on consumption by using an equation, namely:

Consumption expenditure depends on both national income and interest rates, or:

$$C = C(Y, i) \tag{21.2}$$

The marginal propensity to consume out of national income is a positive fraction, $0 < (\Delta C/\Delta Y) < 1$, and the relationship between consumption and interest rates is negative, $\Delta C/\Delta i < 0$.

When consumption expenditure is plotted relative to national income as in the 45° line diagrams of Chapters 18 and 19, a change in the interest rate *shifts the consumption function* but does not change its slope.

Interest rates and investment expenditure

In Chapters 16 and 18 we defined investment expenditure as the purchase of currently produced fixed capital, which includes plants, machinery and equipment; and inventories of raw materials, components, and finished goods. Spending on new residential and non-residential construction is also included in investment. Assume investment is independent of current income and therefore an autonomous component of aggregate expenditure. However, the interest rates determined in money and financial markets affect investment expenditure.

The data in Chapters 16 and 18 showed investment at about 20 percent of GDP in 2019 but with the *level* of investment spending changing from year to year within a range of +/- 11 percent. Although the total change in inventories is quite small, this component of total investment is volatile and contributes to the fluctuations in the total level of investment. Interest rate changes are responsible for some part of the volatility in investment spending.

Government capital expenditures on buildings, roads, bridges, and machinery and equipment are a part of government expenditure G. We treat government capital expenditure as part of fiscal policy and include it in G, not in I.

Businesses spend on fixed capital, plant and equipment to expand their output capacity if they expect growth in demand for their output, or if they see opportunities to reduce costs by adopting new technology and production techniques. Wireless companies like *Bell Canada*, *Rogers* and *Telus* spend continuously on new equipment to accommodate subscriber growth and new products and applications that require more and faster data and voice transmission. Auto makers add to or reduce assembly capacity and develop new product and production technologies to remain competitive and to meet needs for increased fuel efficiency. Solar, wind energy and biofuel companies build new solar farms, wind farms and ethanol plants to provide new sources of electricity and fuels.

The firm's decision to invest is based on its *expectation* of future markets and profits that will justify the estimated cost of new plant, equipment and technologies. Financial markets provide some important guidance.

The current market values of existing firms are the *present values of their expected profits*. A firm thinking about entering an industry or expanding its current capacity can compare the cost of building a new plant and buying equipment with the market value of capital already in the industry. The investment looks profitable if the cost to enter the industry or build and install new capacity is less than the value the market places on existing businesses. Alternatively, if the value the market places on existing business is less than the capital cost of new business there is no incentive to invest in more plant and equipment. However, there might be an opportunity to enter the industry, or expand by taking over an existing business.

The present value of expected profits depends on the interest rate. Changes in interest rates change both the values the market puts on existing businesses and productive capacity and the costs of financing new investment. A rise in interest rates lowers the market value of existing firms and increases the costs of financing new investment. A fall in interest rates increases current market values and lowers financing costs. As a result, investment expenditures are inversely related to interest rates, if all other conditions are constant.

Inventory management is another important part of investment expenditure. Some firms hold inventories of basic inputs to production like raw materials and may also hold components and finished product. Other firms organize their production and coordinate with suppliers to minimize inventories to achieve 'just in time' delivery of inputs. Financial services firms often hold inventories of bonds and other assets to help customers adjust their portfolios.

Inventories can accommodate differences in the timing of production and sales for the benefit of both producers and consumers. If demand for output rises sharply, plant capacity cannot be changed overnight. If demand exceeds current output, sellers would rather not disappoint potential customers. Car dealers hold inventories in part to help smooth the flow of production, and in part to be able to offer immediate delivery. Retail stores carry inventories so customers can buy what they want when they want it. As demand fluctuates, it can be more efficient to allow inventories of finished goods to fluctuate than to try to adjust production to volatile market conditions.

But inventories involve costs. To the producer, unsold goods represent costs of labour, materials, and energy paid but not yet recovered from the sale of the product. These costs have to be financed, either by borrowing or tying up internal funds. Retailers have similar carrying costs for their inventories. Thus, interest rates determine the important finance costs of holding inventories. If we assume prices are constant and interest rates rise, producers and retailers will want smaller inventories. Alternatively, if prices are rising, the difference between the nominal interest rate and the rate of inflation is the real cost of carrying inventories.

The **investment function** is based on these explanations of expenditure on fixed capital and inventories. The negative effect of interest rates in the investment function, $(\Delta I/\Delta i) < 0$, shows that higher interest rates cause lower levels of planned investment expenditure. But how sensitive are investment plans to financing costs? If these financing costs were not a large factor in the investment decision, $\Delta I/\Delta i$ would be small. A rise in the interest rate from i_0 to i_1 would still lower planned investment, but by only a small amount. Alternatively, a larger value for $\Delta I/\Delta i$ would mean that investment plans are sensitive to interest rates.

Investment expenditure depends on the interest rates:

$$I = I(i) \tag{21.3}$$

A rise in the interest rate lowers investment expenditure: $\Delta I/\Delta i < 0$.

> **Investment function,** $I = I(i)$: explains the level of planned investment expenditure at each interest rate.

When plotted in a diagram with interest rate (i) on the vertical axis and investment (I) on the horizontal axis, the slope of the investment function, $I = I(i)$, is $-(\Delta i/\Delta I)$. The *position* of the investment function reflects the effect of all factors, other than interest rates, that affect investment decisions. The price of new capital equipment, optimism or pessimism about future markets and market growth, the introduction of new technologies embodied in newly available equipment, and many other factors underlie investment decisions. Changes in any of these conditions would *shift the I function* and change planned investment at every interest rate. The sharp drop in oil company expenditures on new equipment and production capacity in response to the collapse in oil prices is a clear example of a shift in the investment function. Increased business confidence and expectations of stronger and larger markets shift the *I* curve to the right. Pessimism shifts it to the left.

COVID-19 business shutdowns, for example, and the sharp drop in crude oil prices will dramatically reduce business spending on new buildings, equipment and resource development.

The volatility of investment that causes business cycle fluctuations in output and national income comes from volatility in business profit expectations, rather than from interest rates. Changes in investment, a result of changes in interest rates or as a result of other factors, shift aggregate expenditure and work through the multiplier to change AD, output, and employment. The reaction of investment expenditure to changes in interest rates provides the important link in the monetary transmission mechanism but does not explain the volatility of investment expenditure we saw in Chapter 18.

Exchange rates and net exports

The changes in foreign exchange rates caused by changes in interest rates affect the competitiveness and profitability of imports and exports relative to domestically produced goods and services. A rise in interest rates leads to an appreciation of the domestic currency. Import prices fall relative to the prices of domestic goods and services. Exports become less competitive and less profitable. Imports rise and exports fall, lowering the net export component of aggregate expenditure and demand. Alternatively, a fall in interest rates leads to a depreciation of the domestic currency. Prices of imported goods and services rise relative to the prices of domestic goods and services. Exports are more competitive and more profitable. Net exports increase.

In Chapter 18 we assumed exports were autonomous, independent of national income but dependent on foreign incomes, foreign prices relative to domestic prices, and the exchange rate, which we held constant. Imports were a function of national income, based on a marginal propensity to import, with an autonomous component to capture relative price and exchange rate conditions. Exchange rates were assumed to be constant.

Dropping the assumption that the exchange rate is constant makes the important third link between interest rates and aggregate expenditure through net exports. Exchange rate effects reinforce the negative relationship between interest rates and expenditures in the consumption and investment functions. If interest rates rise, other things constant, the domestic currency appreciates and the exchange rate, *er*, falls. Exports fall, and imports rise, reducing net exports and aggregate expenditure. A net export function that describes this relationship would be:

$$NX = NX[er(i), Y, \dots] \tag{21.4}$$

In Equation 21.4, the variable $er(i)$ captures the effect of interest rates on exchange rates, and exchange rates on net exports. The variable Y captures the effect of changes in Y through the marginal propensity to import. From the foreign exchange market we know that a rise in interest rates leads to an appreciation of the domestic currency that lowers the exchange rate, $(\Delta er/\Delta i) < 0$. Also, a fall in the exchange rate lowers net exports, $(\Delta NX/\Delta er) > 0$.

The appreciation of the Canadian dollar that reduced the Canadian/US dollar exchange rate from \$1.57Cdn for \$1.00US in 2002 to \$1.014Cdn to \$1.00US in March 2008 and \$0.9814Cdn to \$1.00US in November 2012 illustrates the point. Although due more to the rise in commodity and energy prices than to interest rate differentials, the lower exchange rate increased imports and reduced the viability of manufacturing based on exports to the US market, or competition with imports. To the extent that interest rate changes affect exchange rates, they also change net exports and aggregate expenditure.

The depreciation of the Canadian dollar following the collapse of energy and commodity prices in late 2014, and the subsequent lowering of interest rates by the Bank of Canada raised the Cdn/US dollar exchange rate to $1.31Cdn to $1.00US in early August 2015. Although only partly a result of lower Canadian interest rates, this lower dollar makes exports more competitive and profitable and imports more expensive. Over time, net exports should increase and increase aggregate demand.

Figure 21.5: Interest rates & autonomous expenditure

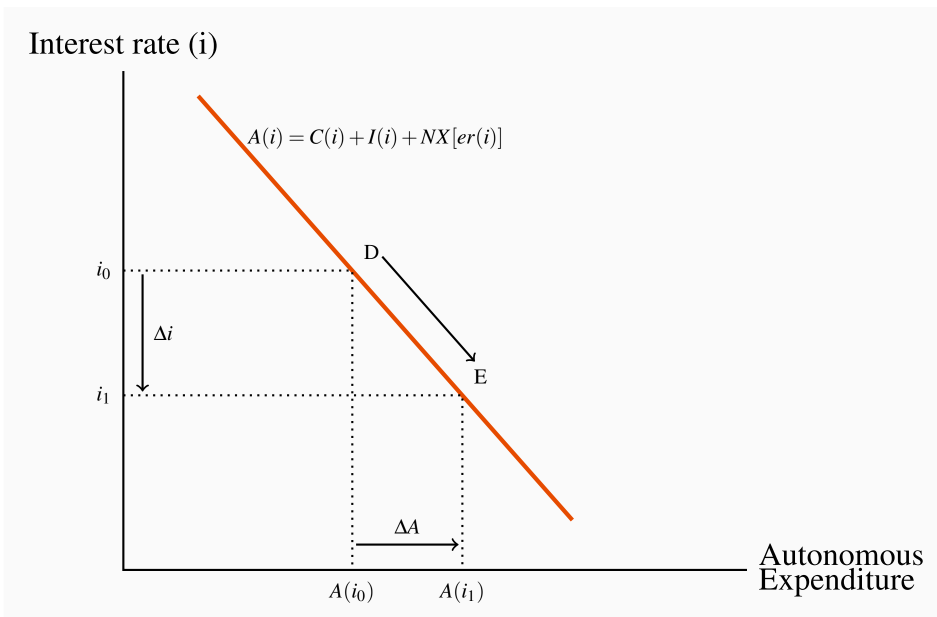


Figure 21.5 summarizes the relationship between interest rates and expenditures, assuming all things other than interest rates and exchange rates are constant. The downward sloping line $A(i)$ illustrates the inverse relationship between the consumption, investment, and net export components of autonomous expenditure and the interest rate. Starting with interest rate i_0, the level expenditure related to interest rates is $A(i_0)$, given by point D on the expenditure function. A fall in interest rates from i_0 to i_1 increases expenditure to $A(i_1)$, moving along the expenditure function to point E. Lower interest rates increase consumption and investment expenditure directly through wealth and cost and availability of finance effects. Lower interest rates also increase net exports through the effects of lower interest rates on the foreign exchange rate. A rise in interest rates would have the opposite effect.

The changes in interest rates and exchange rates are the key linkages between the monetary and financial sector and aggregate demand.

21.6 The monetary transmission mechanism

We can now summarize and illustrate the relationships that transmit changes in money, financial markets, and interest rates to aggregate demand, output, and employment. There are four linkages in the transmission mechanism:

1. With prices constant, changes in money supply change interest rates.
2. Changes in interest rates change consumption expenditure through the wealth effect and the cost and availability of credit.
3. Changes in interest rates also cause changes in planned investment expenditure through the cost and availability of credit to finance the purchase of capital equipment and to carry inventories.
4. Changes in interest rates also cause changes in exchange rates, which change the price competitiveness and profitability of trade goods and services.

Working through these linkages, the effects of changes in money and interest rates on aggregate expenditure, aggregate demand and equilibrium real GDP is illustrated as follows:

$$\Delta M \rightarrow \Delta i \rightarrow \begin{pmatrix} \Delta C + \Delta I \\ \Delta er \rightarrow \Delta NX \end{pmatrix} \rightarrow \Delta \text{AE} \times \textbf{multiplier} \rightarrow \Delta \text{AD} \rightarrow \Delta Y$$

Figure 21.6: The monetary transmission mechanism

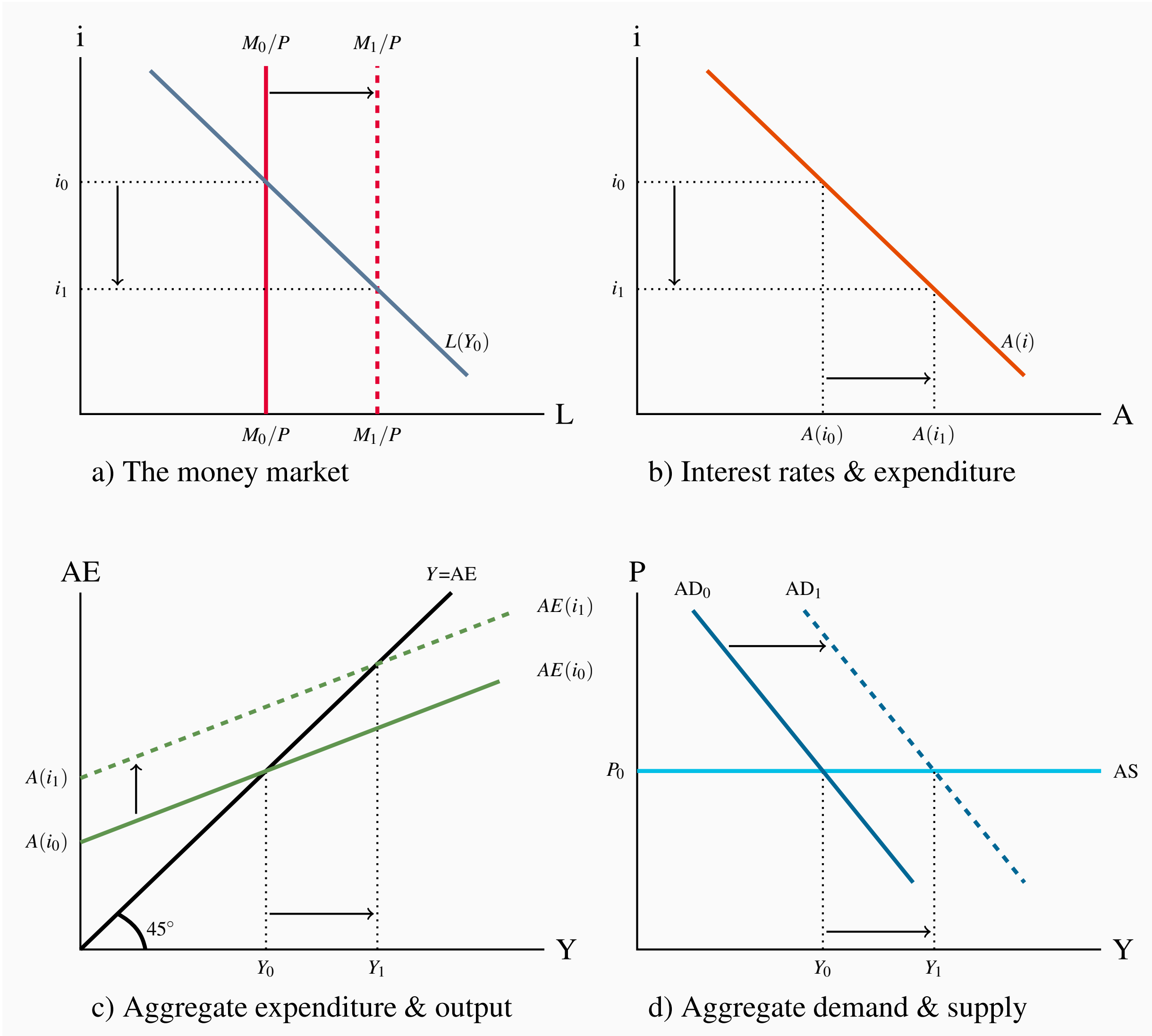


Figure 21.6 shows the transmission mechanism using four interrelated diagrams: a) the money market, b) interest rates and planned expenditure, c) aggregate expenditure and equilibrium output, and d) aggregate demand and supply, output, and prices. We continue to assume a constant price level, as the diagrams show. Changes in the money and financial sector affect aggregate demand and output, to add another dimension to our understanding of the sources of AD and fluctuations in AD.

To see the linkages in the transmission mechanism start in Panel a) with an equilibrium interest rate i_0 determined by the initial money supply M_0/P and demand for money $L(Y_0)$. This interest rate i_0 induces autonomous expenditure $A(i_0)$ in Panel b). That autonomous expenditure is the vertical intercept $A(i_0)$ of the aggregate expenditure function AE(i_0) in Panel c), and through the multiplier the equilibrium GDP, Y_0. In Panel d) the corresponding aggregate demand curve AD_0 crosses the horizontal AS curve at the equilibrium real GDP Y_0.

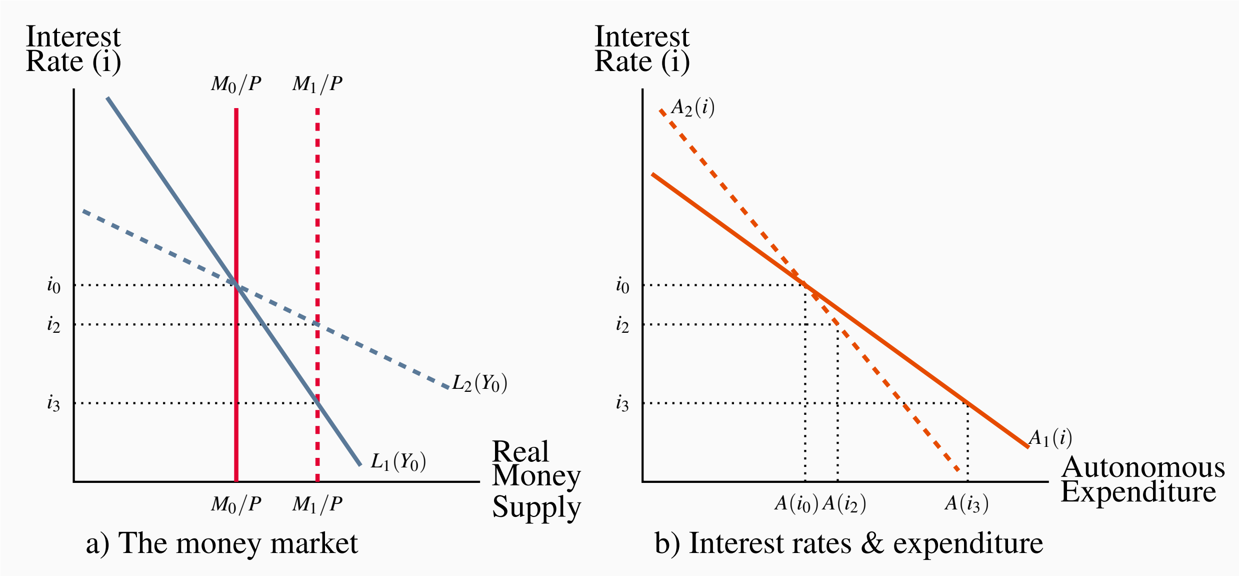
An increase in the money supply in Panel a) lowers equilibrium i to i_1. This causes an increase in autonomous expenditure to $A(i_1)$ in Panel b) and an upward shift in the AE function on Panel c). Increased autonomous expenditure and the multiplier increase equilibrium real GDP and shift the AD curve to the right by the increase in A times the multiplier. The new equilibrium is Y_1 at the price level P_0.

There are several key aspects to these linkages between money, interest rates, and expenditure. The effect of changes in the money supply on interest rates in the money market depends on the slope of the demand curve for real money balances. A steep curve would show that portfolio managers do not react strongly to changes in market interest rates. It would take relatively large changes in rates to get them to change their money balances. Alternatively, if their decisions were very sensitive to the interest rates, the L function would be quite flat. The difference is important to the volatility of financial markets and interest rates, which in turn affect the volatility of expenditure.

Panel a) in Figure 21.7 shows the effects of an increase in money supply under different money demand conditions.

The sensitivity of expenditure to interest rates and financial conditions is a second important aspect of the transmission mechanism. If the interest rate/expenditure function in Panel b) is steep, changes in interest rates will have only small effects on expenditure, aggregate demand, and output. A flatter expenditure function has the opposite implication.

Figure 21.7: The transmission mechanism under different conditions



a) The change in interest rates caused by a change in money supply depends on the interest rate elasticity of the demand for money. L_2 is more elastic than L_1. A small change in interest rates under L_2 induces portfolio managers to increase their money holdings. b) The change in autonomous expenditure caused by a change in interest rates depends on the rate elasticity of autonomous expenditure. Under $A(i_2)$ changes in interest rates have very small effects on expenditures.

Business cycles, output gaps, and policy issues

The effect of money and financial markets on expenditure, output, and employment raises two issues for macroeconomic policy. First, fluctuations in money supply and financial conditions are an important source of business cycle fluctuations in output and employment. These effects are particularly strong and important when the small changes in money supply have big impacts on interest rates and expenditure. A steep $L(i)$ function and a flat expenditure/interest rate function would create these conditions. Stabilization policy would then need to control and stabilize the money supply, a policy approach advocated by Monetarists, who see money supply disturbances as the major source of business cycles. If you can fix money supply at M_0 in Figure 21.6, and the demand for money $L(Y,i)$ and the interest rate/expenditure function are stable, you remove monetary disturbances as a source of business cycles.

This Monetarist approach to money and the financial sector concentrates on the "automatic" stabilization effects of money supply control. With the money supply fixed, any tendency for the economy to experience a recessionary or inflationary gap changes the demand for money, and interest rates change in an offsetting direction. A fall in real output that creates a recessionary gap reduces the demand for money $L(Y,i)$ and, with a fixed money supply, interest rates fall to induce additional expenditure. An inflationary gap would produce an automatic rise in interest rates. The monetary sector automatically resists fluctuations in expenditure and output.

The second policy issue is the alternative to this approach. Discretionary monetary policy would attempt to manage money supply or interest rates or financial conditions more broadly. The objective would be to counter persistent autonomous expenditure and financial disturbances that create output gaps. The intent is to manage aggregate demand in an active way. In other words, if business cycles were caused by shifts and fluctuations in the interest rate/expenditure function in Panel b) of Figure 21.6, monetary policy would react by changing interest rates and money supply and move the economy along the new expenditure function to stabilize autonomous expenditure and aggregate demand. Keynesian and New-Keynesian economists advocate this active approach to policy in the money and financial sector, based on a different and broader view of the sources of business cycles in the economy.

Recent experience extends beyond these two policy concerns. A collapse in the financial sector on the supply side was a major cause of the recession of 2009. Banks and other financial institutions suffered losses on mortgages and other assets followed as energy and commodity prices dropped and expectations of business profits followed. Uncertainty on the part of many lenders about the quality of assets and the risks of lending reduced the availability of credit. Uncertainty on the part of households and businesses reduced their confidence in financial institutions. Although central banks worked to keep interest rates low and bank reserves strong, shifts in the availability of credit and the willingness to borrow shifted the $A(i)$ curve in Figure 21.6 sharply to the left, expenditure fell, and AD shifted left, opening a strong recessionary gap that has been persistent in many industrial countries.

The dramatic shifts in government policies to contain the COVID-19 pandemic changed the scope and scale of monetary and fiscal policies. Mandated business shutdowns caused unprecedented increases in unemployment and losses of business revenues. New monetary policies were developed quickly to fund the dramatic shift in government program expenditures and budget deficits designed to offset support households and businesses. There was a shift, at least during the crisis, from the earlier pattern of monetary policy aimed at an inflation target and fiscal policy constrained by debt ratio targets.

Next

This chapter makes the link between money, interest rates, aggregate demand, and output in the model of the economy. It also shows that monetary policy, working through the monetary transmission mechanism, provides a second policy channel, in addition to fiscal policy, which government might use to stabilize business cycle fluctuations. Chapter 22 studies in detail the monetary policy operations of central banks, including the Bank of Canada.

Example Box 21.1: Bond prices and interest rates

Consider the price of the 4.25 percent bond with a maturity date of June 1, 2018. *Let's assume that the 3 year market rate of interest on the date you buy the bond, say June 1, 2015 is 4.25 percent*. The price of the bond is the present value of the future payments: $4.25 on June 1, 2016, $4.25 on June 1, 2017, and $104.25 on June 1, 2018. Payments to be received two years in the future are discounted twice, and three years in the future three times, to give:

$$PV = (\$4.25)/(1.0425) + (\$4.25)/(1.0425)^2 + (\$104.25)/(1.0425)^3$$
$$PV = \$4.076 + \$3.911 + \$92.013$$
$$PV = \$100.00$$

A bond bought for $100 and held to maturity would *yield* 4.25 percent, the current market rate of interest *assumed in this example*. The bond is trading *at par* because the market price equals the face value.

As an alternative *assume that the 3 year market rate of interest on June 1, 2015 is 1.75 percent*. The price of the bond is then the present value of the future payments: $4.25 on June 1, 2016, $4.25 on June 1, 2017, and $104.25 on June 1, 2018. Payments to be received two years in the future are discounted twice, and three years in the future three times, to give:

$$PV = (\$4.25)/(1.0175) + (\$4.25)/(1.0175)^2 + (\$104.25)/(1.0175)^3$$
$$PV = \$4.177 + \$4.125 + \$98.962$$
$$PV = \$107.26$$

The price of this 4.25 percent bond on June 1, 2018 would be $107.26 per $100 of face value. The assumption that the market rate of interest is 1.75 percent, which is clearly lower than the 4.25 percent coupon on the bond, means the bond trades at a *premium*. The premium price means that buying the bond and holding it to its maturity date will give an annualized return of 1.75 percent on your money. That is the then current assumed 3 year rate.

Taking account of the changes in bond prices as market interest rates change, the **yield on the bond**—the present value of its coupon payment plus the **capital loss** as its price falls to par at maturity—gives a rate of return equal to the market interest rate of 1.75 percent.

Application Box 21.1: A basic guide to financial assets

Three broad classes of financial assets are bought and sold in financial markets. These are *bills*, *bonds*, and *equities*.

Bills are short-term financial assets that make no interest payment to the holder but do make specified cash payment on their maturity date. They trade at a discount. A government treasury bill or T-Bill is an example. Every second week the government sells T-Bills that promise to pay the buyer \$100 for each \$100 of face value on the date that is about three months in the future. The interest earned is the difference between the price paid and the face amount received at the maturity date.

Bonds are longer-term financial assets that pay a fixed money income payment each year and repay their face value on a fixed maturity date. Bonds are marketable, and trade on the bond market between their issue dates and maturity dates at prices determined by supply and demand. As with T-Bills, the return to the holder of a bond depends on the price paid for the bond. In this case, the calculation is more complex however, because it involves a fixed annual money payment and a fixed value at maturity.

Equities are shares in the ownership of the business. They give the holder the right to a share in the profits of the business, either in the form of dividend payments or in terms of the increase in the size of the business if profits are used for business expansion. The shares or stocks in publicly traded businesses can be bought and sold on stock markets like the Toronto Stock Exchange. The financial pages of major newspapers give you daily reports on stocks prices and stock markets. Shareholders' returns from their stock holdings depend on the combination of dividend income they receive and the changes in the market price of the shares they hold. Equity prices are the expected value of the future profits of business. Because expectations of future business performance are volatile, equity prices are volatile and therefore risky. Equities do however offer the prospect of higher long-term returns.

Application Box 21.2: Nominal and real interest rates

Both nominal and real interest rates play important roles in the economy. The **nominal** (or money) **interest rate** is the annual percentage of the principal of a loan that the borrower pays to the lender. It is determined by supply and demand conditions in money markets. The **real interest rate** is the nominal interest rate adjusted for annual changes in the price level (real interest rate = nominal interest rate minus the inflation rate). When the inflation rate is zero, nominal and real interest rates are equal.

Nominal interest rates and financial asset prices are linked. The present value calculation of asset prices uses the nominal rate for discounting. Nominal interest rates and asset prices vary inversely.

Nominal interest rates also affect nominal cash flows of both households and businesses. A rise in nominal rates on lines of credit or mortgages increases the current cash cost of that borrowing. A fall in nominal rates on lines of credit or mortgages releases current cash commitments.

Real interest rates determine the *real* cost of borrowing and the real return to lending.

A family borrows \$200,000 for one year at a nominal interest rate of 5 percent to buy a house. At the end of the year they would owe the lender \$200,000 plus \$10,000 ($\$200,000 \times 0.05$) interest. Their nominal interest cost is \$10,000. If the price level has been constant over the year, their nominal interest cost and their real interest cost are equal at 5 percent.

Suppose however that the all prices are rising by 3 percent a year. The house bought today for \$200,000 will sell for \$206,000 one year from now. Borrowing at 5 percent to buy the house cost \$10,000 but the rise in the price of the house by \$6,000 offsets part of that cost. The real interest cost is $\$10,000 - \$6,000 = \$4,000$. The real interest rate is 2 percent based on the nominal interest rate of 5 percent minus the change in the price level of 3 percent.

With inflation rates greater than zero, lenders' real interest earnings are less than nominal interest earnings. In the preceding example, the mortgage lender's real return was just 2 percent (5 percent – 3 percent inflation) because the \$210,000 received at the end of the year had its purchasing power reduced to approximately \$204,000 by the 3 percent rise in the price level.

Nominal and real interest rates affect expenditure decisions by their effects on asset prices, cash flows, and the real costs and returns involved in borrowing and lending.

Key Concepts

A **financial portfolio** is a mixed holding of money and other financial assets, such as bonds and equities, structured, to balance expected return and risk.

The price of a financial asset like a bond that promises to make future payments is the **present value** of those payments. Because current interest rates are used to discount future payments and determine this present value, **bond prices and interest rates are inversely related**.

The **demand for money** (*L*) is a demand for real money balances measured in terms of purchasing power over goods and services. It arises from the portfolio decisions people make about the form in which to hold their wealth. Holding money reduces the costs of making both routine and unexpected **transactions**. It also provides a **safe asset**, with a fixed nominal price, as a store of wealth. The cost of holding money is the **interest income** and potential capital gain sacrificed by not holding bonds.

The quantity of real money demanded rises with real incomes, to finance higher transactions, and falls with higher nominal interest rates, the opportunity cost of holding money instead of bonds. The demand for money function is $L = kY - hi$.

The **interest rate** (*i*), is determined by supply and demand in the money market, together with supply and demand in the bond market. As people adjust the holdings of bonds and money in their wealth portfolios, bond prices and yields adjust to clear both bond and money markets simultaneously.

Changes in interest rates lead to changes in **foreign exchange rates** that change net exports. The international sector makes an additional link between money, interest rates, and expenditure.

Interest rates play a key role in the **transmission mechanism** that links money and financial markets to aggregate expenditure.

Household consumption expenditure and business investment expenditure are dependent, in part, on interest rates. A higher interest rate reduces household **wealth** and increases the **finance costs** of borrowing. Lower wealth and higher finance costs reduce planned autonomous consumption, shifting the consumption function down. Lower interest rates have the opposite effect.

The **monetary transmission mechanism** links changes in money supply to changes in aggregate expenditure, aggregate demand, and output through interest rates and exchange rates.

Exercises for Chapter 21

Exercise 21.1 If the current market interest rate is 3 percent and a bond promises a coupon of \$3 each year in perpetuity (forever), what is the current market price of the bond? Suppose you were holding such a bond and current market interest rates fell from 3 percent to 2.5 percent. Would you be pleased or disappointed by the return on your bond holding? Why?

Exercise 21.2 Suppose you are holding a bond that will pay \$5 each year for the next two years from today and mature two years from today.

(a) If current two-year market interest rates are 5 percent, what is the market price of your bond?

(b) If market interest rates rise tomorrow to 6 percent, what happens to the market price of your bond?

(c) What is the "market risk" in holding bonds?

Exercise 21.3 Draw a diagram to illustrate the relationship between the demand for real money balances (L), GDP (Y) and the interest rate (i), $L = kY - hi$, when real GDP has a given value Y_0.

(a) Explain your choice of the intersection of your demand for money function with the horizontal axis, and your choice of the slope of the function.

(b) Using your diagram, illustrate and explain the quantity of real money balances demanded for a specific interest rate, say i_0. Pay particular attention to the underlying motives for holding these money balances.

(c) Suppose interest rates declined from your initial assumption of i_0 to a new lower rate i_1. Illustrate and explain the effect of the change in interest rates on the demand for money balances.

(d) Holding interest rates constant at either i_0 or i_1, suppose real GDP were to increase. Illustrate and explain the effect of the increase in real GDP on the demand function and the quantity of real money balances people hold.

Exercise 21.4 Today it costs \$1.25Cdn to buy \$1US. Suppose tomorrow US interest rates rise. What would happen to the foreign exchange rate between Canadian and US dollars? Explain why.

Exercise 21.5

(a) Draw a diagram to illustrate equilibrium in the money market.

(b) Starting from your initial equilibrium, suppose real national income (Y) increased. Illustrate and explain how the money market would adjust to this change in economic conditions.

(c) How does the interest rate in the new equilibrium compare with the interest rate in the initial equilibrium?

Exercise 21.6 Construct a set of diagrams that shows the monetary transmission mechanism linking interest rates to aggregate demand and output. Using these diagrams, show and explain:

(a) How a reduction in the money supply would affect aggregate demand and output.

(b) Alternatively, how an increase in the precautionary demand for money balances caused by terrorist activity, or severe weather events, or an increase in uncertainty in general would affect aggregate demand and output. Assume the money supply is held constant.

(c) Alternatively, how would an increase in autonomous investment expenditure and exports affect aggregate demand, output, and interest rates?

Chapter 22

Central banking and monetary policy

In this chapter we will explore:

On January 21, 2015 Stephen Poloz, the Governor of the Bank of Canada, announced a cut in the Bank's key monetary policy rate, the overnight rate, from 1.00% to 0.75%. The Bank wanted provide monetary stimulus to offset the negative growth rate and inflation rate effects of the sharp drop in oil prices. A further cut in the overnight rate to 0.50% followed on July 15, 2015. The Bank's forecast for growth and inflation in Canada had been revised downward based on weakness in the international economy, particularly the slowdown in growth in China and declining commodity prices. Although growth in real GDP was stronger in 2016 the output gap remained at an estimated -1.25% to -0.25% based on estimation procedures. The Bank of Canada in its April 2017 *Monetary Policy Report*, expecting slow economic growth and stable inflation beyond the first half of 2017, maintained the 0.50% overnight rate.

But economic conditions improved sooner than the Bank of Canada anticipated. Increased global economic growth, and growth in Canadian real GDP by September 2017 and thereafter, called for some reduction in the level of monetary stimulus. What followed was a series of increases in the Bank's overnight interest rate setting first to 1.00% and then, in steps of 0.25%, until the setting of 1.75% in October 2018. Then in January 2019 the Bank held the overnight rate at 1.75%, noting slower growth both globally and in Canada and a fall in the price of oil. The pressure on the inflation rate had eased.

Conditions changed in early 2020. The large losses of employment output and income due to COVID-19 containment forced the Bank of Canada to introduce policy tools beyond its overnight interest rate settings. New government programs required deficit financing, funded by bond issues. 'Quantitative Easing', discussed below, expanded the monetary base to accommodate involved purchases of this government debt.

These are recent examples of monetary policy decisions by the Bank of Canada to defend its *inflation rate target* and support the economy's growth rate. This chapter examines the role of the central bank. The central bank is responsible for monetary policy. Its monopoly control of the supply of cash, or monetary base, gives it a powerful influence in financial markets. Sometimes the central bank controls the monetary base to control the supply of money. Other times it controls short-term interest rates. In either case, central bank actions are designed to affect inflation, output, and employment. They work through the transmission mechanism that links monetary policy to aggregate demand, as we discussed in the last chapter.

22.1 Central banking and the Bank of Canada

Most countries have a **central bank**. Some of these central banks, like the Bank of England, were private firms originally in business for profit, but began to operate in part to promote stability in financial market conditions. The focus of their business shifted to take on an informal role in what is now called monetary policy. As governments also became interested in monetary policy, central banking institutions were established in countries where none previously existed. The Federal Reserve System, the United States central bank, was created under federal law in 1913. It is a system of 12 regional banks, each owned by the commercial banks that are its members. Canada's central bank, the Bank of Canada, was set up and started operations in 1935 as a privately owned institution, but was nationalized in 1938. In the United Kingdom, the previously private owned Bank of England was nationalized in 1947.

> **Central bank**: an institution that conducts monetary policy using its control of monetary base and interest rates.

In every case, the important distinction between a private bank and a central bank is the purpose that drives the institution's operations. Private banks are profit-oriented businesses providing financial services to businesses and households. Central banks conduct their operations to influence the behaviour of other banks and intermediaries in the financial system. Profits are *not* the motive behind central banks' operations, although they do make profits. They also serve as banker to the government and to the banks. But their primary role and responsibility is to conduct **monetary policy**: To control inflation and support economic growth through their control of the monetary base and interest rates, and perhaps the foreign exchange rate.

> **Monetary policy**: central bank action to control inflation and support economic growth through control of the money supply, interest rates, and exchange rates in order to change aggregate demand and economic performance.

The Bank of Canada is Canada's central bank. A visit to its website `www.bankofcanada.ca` provides detailed information on its structure, operations and monetary policy objectives. The information the bank provides on monetary policy is of particular relevance for this chapter.

See `http://www.bankofcanada.ca/core-functions/monetary-policy/`.

The Bank of Canada is led by the Governing Council, the policy-making body of the Bank. The Governing Council is made up of the Governor, the Senior Deputy Governor and four Deputy Governors. The current governor is Stephen Poloz. Currently, the Council's main tool for conducting monetary policy is the target for the overnight rate (also known as the key policy rate). Stephen Poloz, like governors before him, manages the Bank's balance sheet to implement monetary policy. He can expand the Bank's asset holdings and pay for that expansion by creating new Bank of Canada liabilities, which are additions to the monetary base. Alternatively, he can sell some of the Bank's assets, destroying an equal amount of liabilities and monetary base. No *reserve requirements* (explained in Section 22.2) limit these operations. The management of the Bank's balance sheet and the monetary base depends on the wisdom and judgment of the Governor and management of the Bank. They work to get the monetary base and interest rates that are appropriate for the economy.

There is a further interesting difference between the commercial and central bank balance sheets. Private banks concentrate on their deposit base and loan operations. These are the main entries in their balance sheets and the source of their banking profits as discussed in Chapter 20. The Bank of Canada, by contrast, does very little direct lending, and any it does is of very short duration. Indeed lending to banks and other financial institutions, advances to members of the Canadian Payments Association, which would be central bank loans, were $143 million at the end of January 2019 and zero at the end of August 2019.

Nor does the Bank of Canada hold many deposits. It does not need deposits as a source of funds. Deposit facilities are provided to the commercial banks and other members of the Canadian Payments Association for their use in settling cheque-clearing balances and transfers among the banks, and to the Government of Canada. Cheques issued by the Government of Canada, like income tax refunds, Old Age Security payments, and Employment Insurance benefits, are drawn on the government's account in the Bank of Canada. This difference in the structure of operations again shows the difference between profit-oriented commercial banks and a central bank with responsibility for monetary policy.

Having the power to conduct monetary policy is one thing; how you use it is another. The Bank of Canada's responsibilities are set out in the Bank of Canada Act, the act of Parliament that established the Bank in 1934. According to the Act, the Bank is to conduct its policy in ways that support the economy by reducing fluctuations in output, prices, and employment while protecting the external value of the currency. In terms of our study of the economy, we can describe these goals of monetary policy as the pursuit of potential output and low, stable inflation rates.

Exactly how the Bank is to achieve those objectives has been, and continues to be, a topic for discussion and debate. Over the years, our understanding of what monetary policy can and cannot do has evolved, as have the Bank's interpretation of its mandate and the techniques it uses to conduct monetary policy. The Canadian economist Robert Mundell has been a major contributor to this work. His explanations of the transmission mechanism and the strength of monetary policy under different foreign exchange rate systems were recognized by his 1999 Nobel Prize in Economics.

Currently, the Bank works to maintain inflation within a target range of 1 percent to 3 percent, but that has not always been its explicit policy objective. Gordon Thiessen, a recent Governor of the Bank of Canada, provides an interesting overview of the evolution of monetary policy in Canada from the 1930s to the end of the 1990s[1].

22.2 Central bank operating techniques

The money supply—currency in circulation plus the deposits of the commercial or chartered banks—is partly a liability of the central bank (currency) and partly a liability of the commercial banks (deposits). In Chapter 20 we discussed the *monetary base* supplied by the central bank. You will recall that the *money multiplier* ties the size of the money supply to the size of the monetary base. The money multiplier is larger when:

1. the reserve ratio (rr) banks hold is smaller; and
2. the amount of currency the non-bank public wishes to hold is small and constant.

If these two ratios are constant, the central bank can change the size of the money supply by changing the size of the monetary base.

In general, central banks have three main techniques for the control of the monetary base and the money supply. These are:

1. Establishing reserve requirements
2. Using open-market operations
3. Adjusting central bank lending rates

Not all central banks use all three techniques, but we will examine each of them. Later we will see that the Bank of Canada has some additional operating techniques it uses to influence interest rates in the short run.

In the financial crisis and deep recession of 2008-2009, central banks developed additional techniques to support the banking system, the availability of credit, and the money supply. The 'Quantitative Easing' techniques used in the US and more recently in Europe are examples of these techniques, as is the Bank of Canada's policy response to the COVID-19 crisis.

Reserve requirements

In some cases, commercial banks operate under a legal **required reserve ratio**. They are required by law to hold cash reserves and central bank deposits not less than some specified percentage of their deposit liabilities.

[1] Thiessen, G., "Can a Bank Change?" *Bank of Canada Review*, Winter 2000/2001, pp. 35-46, and also available at `http://www.bankofcanada.ca/2000/10/can-a-bank-change/`.

Required reserve ratio: a legal minimum ratio of cash reserves to deposits.

Banks can hold more than the required reserves but not less. If their reserves fall below the required amount, they must borrow cash, from the central bank, to restore their required reserve ratio. Since a loan from the central bank carries an interest rate, usually higher than the market interest rate, borrowing imposes a cost on the bank and lowers profitability. Banks usually hold slightly larger reserves than required to avoid the costs of falling short.

A required reserve ratio is essentially a regulation used to give the central bank control of the money supply. The reserve ratio is a key determinant of the money multiplier. If a central bank has the power to change the commercial banks' required reserve ratio, it can use it to change the money supply. For a given monetary base, a rise in the required reserve ratio reduces the size of the money multiplier and the money supply. A reduction in the reserve ratio has the opposite effect.

However, changing required reserve ratios are blunt techniques for monetary control. They simultaneously affect the reserve positions of all banks in a system and require large adjustments in financial markets. As a result, changes in reserve ratios are not widely used as techniques for money supply control.

Required reserve ratios are different in different national banking systems. In the United States, for example, the Federal Reserve is authorized to impose reserve requirements of 8 percent to 14 percent on chequable deposits, and up to 9 percent on non-personal time deposits. As of February 2002, the ratios were set at 10 percent for chequable deposits and 0 percent for time deposits. The European Central Bank also imposes reserve requirements. In 2012, both India and China reduced deposit reserve ratios on several occasions to encourage monetary expansion in the face of declining GDP growth rates.

Until 1994, banks in Canada were subject to legal minimum reserve requirements. These have now been phased out, as have reserve requirements in many other countries. In Canada, the banks hold reserves made up of very small settlement balances in the Bank of Canada, in addition to their cash holdings. The banks decide the size of their reserve ratios based on their own assessments of their reserve needs, rather than a legal requirement. We will see later that reserve holdings, and the Bank of Canada's management of the available cash reserves, are important to the implementation of monetary policy in Canada.

The absence of legal reserve requirements in Canada means that reserve ratios in the banking system change from time to time. They may change as the banks change their outlook on financial conditions and their evaluation of banker's risk. These changes are linked to the profit motive of the banks rather than the control interests of the central bank. Whether they come from central bank action or commercial bank asset management, changes in the banks' reserve ratio change the money multiplier and the money supply.

Open market operations

Open market operations are central bank purchases or sales of government securities in the open financial market. They are the main technique used by central banks to manage the *size of the monetary base*. Whereas reserve requirements affect the money supply through control of the money multiplier, open market operations work directly on the monetary base. Since the money supply is the monetary base multiplied by the money multiplier, open market operations alter the money supply.

> **Open market operation**: central bank purchases or sales of government securities in the open financial market.

Central banks use open market operations to provide the monetary base needed to support the demand for money and the increase in the demand for money as the economy grows. If monetary policy is conducted by setting interest rates, as discussed later in the chapter, open market operations are passive. They provide the monetary base needed to meet the demand for money at the interest rate set by the central bank. An open market purchase makes a permanent addition to the central bank's assets and monetary base.

There are times when monetary policy is conducted through control of the money supply. If the money multiplier is constant, a central bank can control the size of the money supply by controlling the monetary base using open market operations. Open market purchases increase the monetary base and increased bank lending increases the money supply. Open market sales have the opposite effect.

In times of financial and economic crisis, as in 2008 and 2009, open market operations are used along with interest rate setting. High uncertainty in financial markets and falling demand in goods-and-services markets increased the demand for liquid cash balances. If interest rates are reduced close to zero without increasing lending and spending and asset demand, the central bank may undertake "quantitative easing," using open market purchase to increase the monetary base and offset a shortage of liquidity in the economy. This topic comes up again after we look at monetary policy in more normal times.

Table 22.1 illustrates an open market purchase and its effect on bank reserves and the money supply. To keep the example simple, we will assume the banks hold reserves equal to 5 percent of their deposits, $rr = 0.05$, but the public's currency holdings are constant. This means a simple money multiplier is equal to $1/rr = 1/0.05 = 20$.

Table 22.1: An open market purchase and the money supply

Central Bank: Assets	Central Bank: Liabilities	Commercial Banks: Assets	Commercial Banks: Liabilities
1. Open market purchase of $100 million in government bonds.			
Gov't bond +100	Cheque issued +100	No change	No change
2. Pension fund deposits proceeds of bond sale in commercial bank.			
No change	No change	Central bank cheque +100	Pension fund deposit account +100
3. Central bank cheque clears giving commercial banks $100 million in cash.			
No change	Cheque o/s −100 Cash issued +100	Central bank cheque −100 Cash reserves +100 (excess reserves +95)	No change
4. Commercial banks increase lending and create new deposits backed by their increased cash reserves.			
No change	No change	Loans +1,900	Deposits +1,900
5. Final effect of central bank open market purchase.			
Gov't bond +100	Cash issued +100 (ΔMB)	Cash reserves +100 Loans +1,900 +2,000	Deposits +2,000

In the example, the central bank buys $100 million of government bonds on the open market. Assume a large pension fund sold these bonds, and received in payment a cheque for $100 million issued by the central bank. This transaction is recorded (as $100) under item 1 in the table.

Item 2 in the table records the pension fund's deposit of the central bank cheque in the commercial banking system. The commercial bank issues a deposit to the pension fund in return for the cheque drawn on the central bank.

The commercial bank does not want to hold the central bank cheque. It presents it for payment and receives, in this example, cash in the form of central bank notes. Cash is a reserve asset for the commercial bank. In item 3 in the table, the central bank has created new monetary base, which has increased the cash reserves of the commercial bank by \$100. The commercial bank now has new reserves of \$100 against its increased deposit liabilities of \$100. Based on its reserve ratio $rr = 0.05$, it has *excess reserves* of \$95.

Excess reserves in the commercial banking system support an increase in lending and the creation of new bank deposits. Item 4 in the table shows the final results of this loan and deposit expansion, for the entire banking system. Based on a simple money multiplier of 20, we know that the increase in the monetary base in the form of new cash reserves by \$100 will result in an increase in the money supply of \$2,000. Bank lending and deposit creation continue until total deposits have increased by \$2,000, based on an initial deposit of \$100 and increased lending of \$1,900. Item 5 in the table shows these final results.

In this example, an open-market purchase increased the monetary base and the money supply. The purchase was paid for by the creation of new monetary base. An *open market sale* would have the opposite effect. The monetary base and the money supply would be reduced. An open market operation is a technique a central bank can use to shift the money supply function and affect equilibrium conditions in the money market.

Open market operations are today the principal channel by which central banks, including the Bank of Canada, manage the *longer-term* growth of the monetary base.

Application Box 22.1: Monetary policies to limit economic impact of COVID-19 containment policies

The Bank of Canada took extensive policy action to ease credit, support government programs and help moderate economic impact of business shutdowns and physical distancing, In addition to cutting its overnight interest rate target by 1.5 percentage points to 0.25 percent, the lower effective bound, it conducted lending operations and asset purchases on a unprecedented scale. These included open market purchases of federal government bonds, weekly, and increased purchases of treasury bills. The Bank also developed a new program to purchase provincial bonds and a limited amount of high-grade corporate bonds in the secondary market.

These monetary policy changes supported the financing of the federal government public debt reported earlier in Table 19.4. The Bank of Canada's balance sheet shows that total assets, essentially monetary base, doubled in size to ease credit conditions and accommodate the scale of borrowing by governments needed to fund income support for individuals and grants and loans to support businesses. This monetary accommodation was also designed to create conditions favourable to economic recovery when limits on business, individual and household activities are lifted.

The Bank of Canada Balance Sheet (millions of dollars)

Assets	**25Mar2020**	**22April2020**
Government of Canada direct and guaranteed securities		
Treasury bills	25,773	37,977
Total Canada bonds	76,960	97,673
Mortgage bonds, Provincial money market	16,646	38,741
Total assets	**154,094**	**348,520**
Liabilities and capital		
Notes in circulation	91,538	93,235
Canadian dollar deposits		
Government of Canada	29,520	46,929
Members of Payments Canada	20,619	197,251
Other deposits	10,777	9,250
All other liabilities and capital	1,640	1,856
Total liabilities and capital	154,094	348,520

Source: `https://www.bankofcanada.ca/rates/banking-and-financial-statistics/`

The bank rate

The **bank rate** is the interest rate the central bank charges the commercial banks if the commercial banks borrow reserves. The bank rate or lending rate is set by central banks as a part of their monetary policy operations.

> **Bank rate**: the interest rate the central bank charges on its loans to commercial banks.

Suppose the banks think the minimum safe ratio of reserves to deposits is 5 percent. It does not matter whether this figure is a commercial judgment, as in Canada, or a legal requirement, as in the United States. Banks may also hold a little extra cash to cover day-to-day ups and downs in deposits and withdrawals, but maximum profit requires minimum cash holdings.

One way in which an individual bank can cover a shortage in its reserves is to borrow from other banks that have unexpected excess reserves. This creates a market for monetary base. In Canada, this borrowing and lending takes place on an overnight basis—you borrow today and repay tomorrow, at the overnight interest rate. In the United States, the rate for similar lending and borrowing among banks is the federal funds rate.

If it happens that no other bank in the system has excess reserves to lend, a bank that is short of reserves borrows from the central bank. The interest rate charged is the bank rate, which is set higher than the overnight rate by the central bank, to encourage banks to borrow and lend reserves in the overnight market.

The bank rate is used in different ways by different central banks. There is a long tradition of using changes in the rate as a signal of changes in monetary policy. A cut in the bank rate signals the central bank's intention to increase the monetary base. A rise in the bank rate signals tighter monetary conditions. The role the bank rate currently plays in Canada is discussed later in this chapter.

Money supply versus interest rates

Control of the monetary base through open-market operations and stable desired reserve and cash ratios for the banks and the public give the central bank control of the money supply. This is easy in theory but not in practice.

There are several problems. Can the central bank control the monetary base precisely? The commercial banks can borrow from the central bank at the bank rate when they are short of reserves. Borrowings increase the monetary base. In more difficult financial market circumstances, like those of 2007 to 2009, orderly financial markets may call for large changes in the monetary base to offset extraordinary demands for cash. Meeting these demands takes time and adds to turmoil in markets.

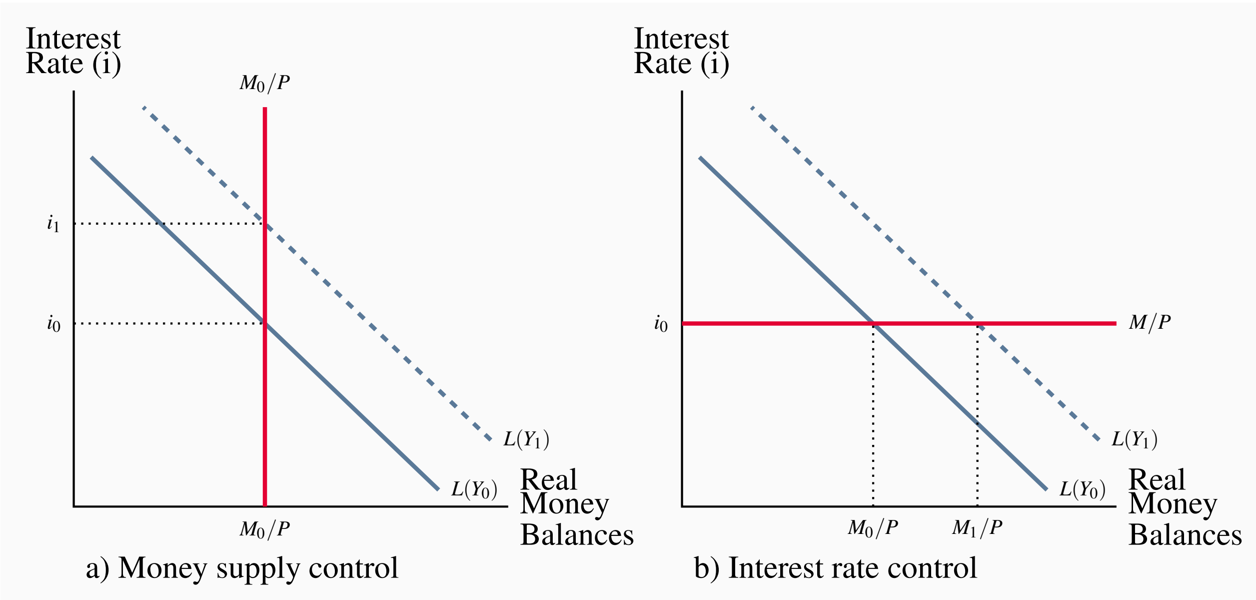
What is the size of the money multiplier? Are desired reserve ratios and cash ratios *stable and predictable* or do they fluctuate? If they fluctuate, the size of the money multiplier is difficult to predict. The money supply function may be unstable.

What money supply measure should the central bank control: MB, M1B+, M2, M2+, or some other aggregate? Households and businesses can shift among the different deposits with different terms and interest rates. Furthermore, the banks are imaginative and competitive in developing new types of deposits.

In short, precise control of the money supply is difficult. Most central banks no longer try. Instead, they set interest rates. The television news and financial press report decisions by the central bank about interest rates, not decisions about money supply. The Bank of Canada and the United States Federal Reserve make regular announcements about their settings of the overnight rate and the federal funds rate, respectively. Several examples of changes in the overnight interest rate by the Bank of Canada were noted in the introduction to this chapter.

Figure 22.1 shows the money market under two different conditions. In both cases we draw the demand for money function $L(Y_0)$ for a given level of real GDP. If the central bank can control money supply, then, for a given level of prices, it fixes the money supply at M_0/P. The equilibrium interest rate is i_0. This is the case in Panel a) of the figure.

Figure 22.1: Money supply control vs. interest rate control



Panel a) assumes the central bank can fix the money supply at M_0/P and the equilibrium rate is i_0. An increase in Y increases demand for money from $L(Y_0)$ to $L(Y_1)$. With a fixed money supply, the interest rate rises to i_1. Alternatively, if the central bank knows the demand for money it can control money supply using interest rates. If Y increased it would increase i to i_1 to reduce the demand for money to its money supply target. A fall in Y would call for a fall in interest rates to control money supply. Panel b) assumes the Bank sets the interest rate at i_0. To do this it must supply whatever quantity of money is demanded at i_0. An increase in Y increases L and results in an increase in M/P. Now the money supply is demand determined. In equilibrium, the central bank supplies exactly the quantity of money demanded at interest rate i_0. The quantity of money supplied is M_1/P. The money supply function is horizontal at the interest rate i_0.

Alternatively, the central bank can fix the interest rate at i_0 and supply whatever money is needed to clear the market at this rate. This is the case in Panel b).

The central bank can fix either the money supply or the interest rate ***but not both****. If it fixes the money supply, it must accept the equilibrium interest rate implied by the demand for money. If it fixes the interest rate, it must accept the equilibrium money supply implied by the demand for money equation. Central banks now do the latter.*

22.3 Monetary policy targets and instruments

A central bank can use the power it has over the monetary base and interest rates to pursue any one of three possible *instrument targets*. It might:

1. Control the foreign exchange rate, or
2. Control the money supply, or
3. Control the inflation rate.

However, *it must choose*. Controlling one of these instrument targets uses all the central bank's power, and it cannot pursue a second target at the same time.

The central bank chooses among these instrument targets based on its judgment as to which target will achieve the best results in terms of its broad *monetary policy objective*: To promote economic stability at potential output with low inflation. The Bank of Canada has conducted its monetary policy in terms of each of these instruments at different times in the recent past.

The brief discussion of the foreign exchange rate in Chapter 21 explained that changes in interest rates will result in changes in the foreign exchange rate. Wealth holders shift their financial portfolios between assets of different countries based on differences in interest rates and bond yields between countries. Rather than allow private supply and demand in the foreign exchange market to set the exchange rate, the central bank can intervene to control the rate. It buys or sells foreign exchange in the market, which affects the supply or demand for Canadian dollars in the foreign exchange market and changes the exchange rate.

Central bank purchases or sales in the foreign exchange market change the domestic monetary base just like open market operations in domestic money market. The domestic money supply and interest rates change until the difference between domestic and foreign interest rates is eliminated. To maintain a fixed **exchange rate target**, the central bank matches domestic interest rates to those set in the country to which it wishes to fix its exchange rate. In Canada, for example, to fix the exchange rate between the Canadian dollar and the US dollar, the Bank of Canada would set its interest rate equal to that set by the Federal Reserve.[2]

> **Exchange rate target**: monetary policy maintains a fixed price for foreign currency in terms of domestic currency.

A central bank may choose to fix the exchange rate because it believes that is the best way to achieve the broader objectives of monetary policy. Canada operated with fixed exchange rates from 1962 to 1970. The Canadian-dollar price of the US dollar was fixed at $1.075, and the Bank of Canada focused its monetary policy on that target. During the late 1950s and early 1960s, there was an intense debate over the monetary policy pursued by the Bank of Canada. Economic

[2]Different exchange rate policies are explained in detail in Chapter 24.

growth was slow, unemployment rates were high, and there was turmoil in financial markets. A fixed exchange rate was seen as the best solution to these economic problems. It essentially gave Canada the monetary policy of the United States, where economic performance had been stronger and more stable than in Canada. The fixed exchange rate target determined Canadian interest rates and money supply until 1970.

Rather than fixing the exchange rate, a central bank can choose to fix the size or growth rate of the domestic money supply. Chapter 21 discussed the money market in terms of a fixed money supply. However, central banks have found that the money multiplier, based on the desired reserve ratios of the banks and cash holdings of the public, is not stable enough to give them control of the money supply directly through their control of the monetary base. Instead, they set interest rates to get their target money supply from the demand for money. If money supply is above their target, they raise interest rates, reducing the demand for money until money holdings fall within their target range. They reduce interest rates if money holdings are less than their target.

In the 1970s, a sharp rise in inflation shifted the focus of monetary policy toward inflation control. At that time, developments in economic theory emphasized a strong link between money, money supply growth, and inflation. Central banks in many industrial countries shifted their focus to money supply control. Canada had dropped the fixed exchange rate target in 1970. In 1975, the Bank of Canada adopted **money supply targets** as its policy instrument and used them until 1982 in an attempt to control inflation and promote a strong economy. By adjusting interest rates based on its understanding of the demand for money balances, the Bank was able to meet the targets for the growth in the money supply M1 that it set and revised from time to time.

> **Money supply target**: a central bank adjusts interest rates and the monetary base to control the nominal money supply, or the rate of growth of the nominal money supply.

However, controlling the money supply required wide fluctuations in interest rates and in the exchange rate. Financial markets did not like this volatility. More importantly, success in controlling M1 did not bring success in controlling inflation. The relationship between money supply, prices, and inflation turned out to be less stable than expected. The Bank abandoned its M1 control targets in 1982 and began a search for a better target for monetary policy.

In the early 1990s, the Bank of Canada and central banks in many other countries, including Australia, New Zealand, Sweden, the United Kingdom, and the European Union, decided to set explicit **inflation rate targets** for monetary policy. The Bank of Canada began to use *interest rate setting* as its **monetary policy instrument**, making changes in the interest rate, as necessary, to keep the Canadian inflation rate within a target range of 1 percent to 3 percent.

The shifts to formal inflation targets for monetary policy in 1991, and the adjustment to that policy shift, were sources of a substantial policy debate in Canada. Inflation was reduced as planned, but with a deep and prolonged recession in real GDP and persistently high rates of unemployment. Sustained economic growth did not resume until the mid-1990s, and by most estimates, including those by the Bank of Canada, the recessionary GDP gap persisted until the end of the decade.[3] The Bank continues today to focus on an inflation rate in the 1 percent to 3 percent range as its monetary policy target. A summary of the costs and benefits of inflation that lie behind its inflation targeting are offered by the Bank of Canada at:

www.bankofcanada.ca/wp-content/uploads/2010/11/benefits_low_inflation.pdf

> **Inflation rate target**: monetary policy objective defined as an announced target inflation rate.
>
> **Monetary policy instrument**: the monetary variable the central bank manipulates in pursuit of its policy target.

Bank of Canada operating techniques

In Canada, the **overnight rate** is now the Bank of Canada's key policy instrument. This is the interest rate that large financial institutions receive or pay on loans from one day until the next. The Bank implements monetary policy by setting a target for the overnight rate at the midpoint of an *operating band* that is plus or minus one-quarter of one percentage point, or 25 basis points, from the target rate.

> **Overnight rate**: the interest rate large financial institutions receive or pay on loans from one day until the next.

The *bank rate* now marks the upper end of this operating band for the overnight rate. It is still the rate at which the Bank of Canada is willing to lend to the banks. The lower end of the operating band, *the deposit rate*, is the interest rate the Bank of Canada pays on deposits. Because the highest cost of borrowing cash is the bank rate, and the lowest return from lending cash is the deposit rate paid by the Bank of Canada, the rate on overnight borrowing and lending among the banks falls within the target range set by the Bank of Canada.

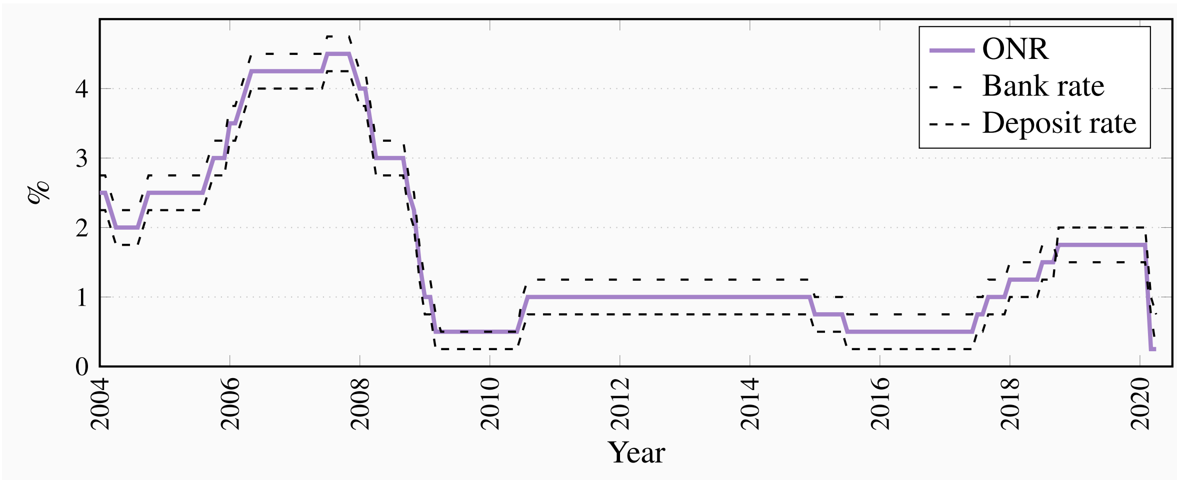
The Bank of Canada tells financial markets the direction in which it wants interest rates to move, by making changes to its target overnight rate and operating band as it did in January, July and October 2018. Changes in the target overnight rate led to changes in the interest rates banks offer to lenders and depositors. A lower target lowers bank lending rates, encouraging more borrowing by households and businesses and a corresponding expansion in the money supply.

Figure 22.2 shows the Bank's settings and changes of the overnight interest rate operating band over the past 15 years. From mid-2004 to late 2007, for example, the plot shows the Bank raised

[3] You can read about this debate in: D. Laidler, and Wm. Robson, The Great Canadian Disinflation: The Economics and Politics of Monetary Policy in Canada 1988-1993 (Toronto: C.D. Howe Institute, 1993). P. Fortin, "The Great Canadian Slump," *Canadian Journal of Economics* 29, No. 4 (August 1996), pp. 1082-1092. C. Freedman and T. Macklem, "A Comment on the 'Great Canadian Slump' ", *Canadian Journal of Economics* 31, No. 3 (August 1998), pp. 646-665.

its overnight rate setting by steps of 0.25 percent (25 basis points) from 2.5 percent to 4.5 percent. Its intention was to reduce monetary stimulus to keep the economy working at potential output, with projected growth of real GDP at 3.0 percent and projected inflation at its target 2.0 percent.

Figure 22.2: The Bank of Canada Operating Band for the Overnight Rate, 2004–2020



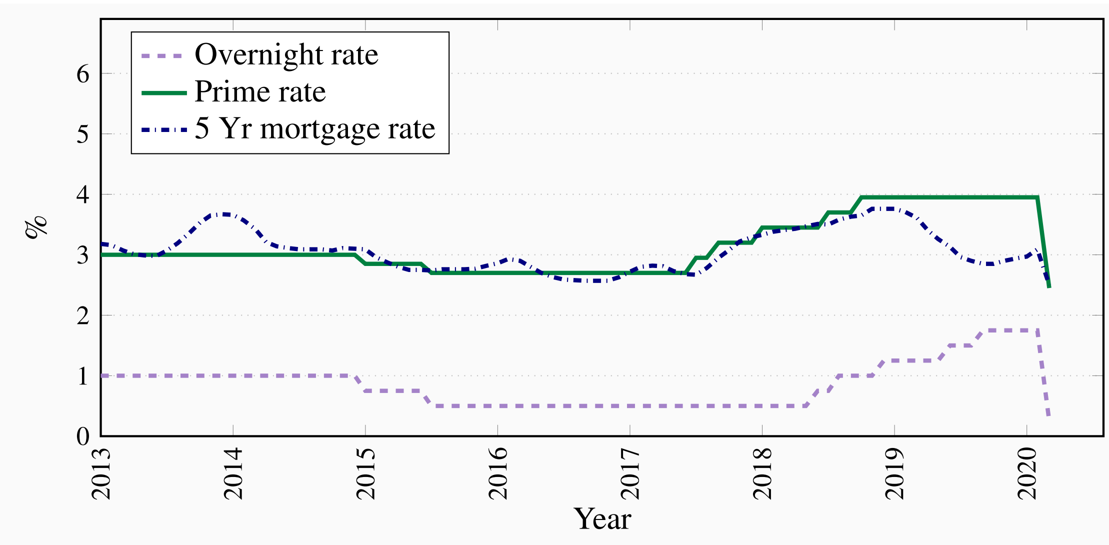
Source: Statistics Canada Table 10-10-0139-01

By contrast from July 2007 to March 2009 the Bank cut its setting for the overnight rate, in steps of 25 to 75 basis points, by a total of 400 basis points from 4.5 percent to 0.5 percent. Its intention then was to provide stimulus to and support for domestic financial markets and economic activity as the global financial crisis developed and spread. As economic conditions improved in 2010 and 2011 the Bank raised the overnight rate to 1.0 percent only to lower it again in 2015 to 0.5 percent, where it remained until mid-2017. Then a series of increases raised the overnight rate to 1.75% in October 2018 and held that rate at least until March 2020 and the outbreak of the COVID-19 epidemic forced a cut in the rate to the effective lower bound of 0.25 percent, followed by quantitative easing in April to support emergency credit needs. Once the crisis passes the Bank can be expected to return to its earlier inflation target and overnight rate approach. You can read the details of the Bank's shift in policy in its press release: `https://www.bankofcanada.ca/2020/04/fad-press-release-2020-04-15/`. You can read a brief analysis of the earlier Bank decisions to change or hold constant its overnight rate target in the press releases on the Bank's website: `http://www.bankofcanada.ca/publications-research/press-releases/`.

By setting the overnight interest rate, the Bank has a direct impact on interest rates that are important to the monetary transmission mechanism. Banks respond to a rise in the overnight rate by raising their **prime lending rate**, which is the base for most of the interest rates on their lending. Rates on business and consumer lines of credit, for example, are linked to the prime rate, and move up and down with it. The connection to mortgage rates is also strong and they move with the overnight rate, although the link is not quite as tight. These interest rates cover about two thirds of bank financing in Canada. They make a strong link between the Bank of Canada's monetary policy action, expenditure, and aggregate demand. Figure 22.3 shows the relationship between changes in the setting of the overnight rate and other interest rates.

Prime lending rate: the base for setting the interest rates charged by banks on loans and lines of credit.

Figure 22.3: The overnight rate, prime rate and 5 year mortgage rate



Source: Statistics Canada Table 10-10-01221, 10-10-0122-01, and Bank of Canada, Banking and Financial Statistics, Interest Rates.

Other countries implement monetary policy by setting similar interest rates. The overnight rate set by the Bank of Canada is comparable to the United States Federal Reserve's target for the 'federal funds rate', and the Bank of England's two-week 'repo rate', and the minimum bid rate for refinancing operations, the 'repo rate', set by the European Central Bank.

Institutional arrangements are the key to the Bank of Canada's use of the overnight rate as its policy instrument. The payments made by individuals and businesses, and their receipts, flow through the banking system. Some are small paper-based transactions that involve the writing of cheques on deposit accounts. Others, and indeed the majority, are transfers of large deposits between bank customers. On any day, an individual bank may take in more deposits through these transfers than it pays out, or pay out more than it takes in. Any difference in either case is settled using balances held on deposit at the Bank of Canada. Technology now allows for *same day settlement* of large-value transactions, and an individual bank's settlement balances can change quickly.

Chartered banks in Canada operate under a *zero settlement balance requirement*. This means that the balances they hold in their deposit accounts at the Bank of Canada cannot be less than zero at the end of the day. If a bank's account is overdrawn from making payments to other banks, it must borrow to cover the overdraft either from another bank or from the Bank of Canada. Borrowing from other banks costs the *overnight rate*. Borrowing from the Bank of Canada costs the *bank rate*, which is set by the Bank of Canada as one-quarter of a percentage point above the overnight rate. As a result, falling short of the zero-balance requirement imposes a cost on a bank, reducing its profitability, which it would like to avoid.

A positive settlement balance also imposes a cost. The Bank of Canada does pay interest on a positive balance in a bank's account, but it pays at its *deposit rate*. That rate is set one-quarter of one percentage point below the overnight rate. Not lending a positive balance to another bank at the overnight rate and accepting the Bank of Canada's *deposit rate* carries an opportunity cost a bank would prefer to avoid. This is a further incentive to maintain a zero settlement balance at the Bank of Canada.

This regulatory and institutional environment gives the Bank of Canada a framework for setting the interest rate to implement its monetary policy. The Bank makes eight scheduled announcements per year about its target for the overnight rate, and the operating band it is setting for the overnight rate. These announcements are made in press releases and include a brief explanation of the economic conditions on which the Bank's rate-setting decision is based.

Special purchases and sales (SPRAs and SRAs)

To maintain the overnight interest rate within the target band, the Bank of Canada must intervene in the market to cover any shortages or remove any surpluses of funds that would push rates beyond its target. The Bank has two tools it uses for this purpose.

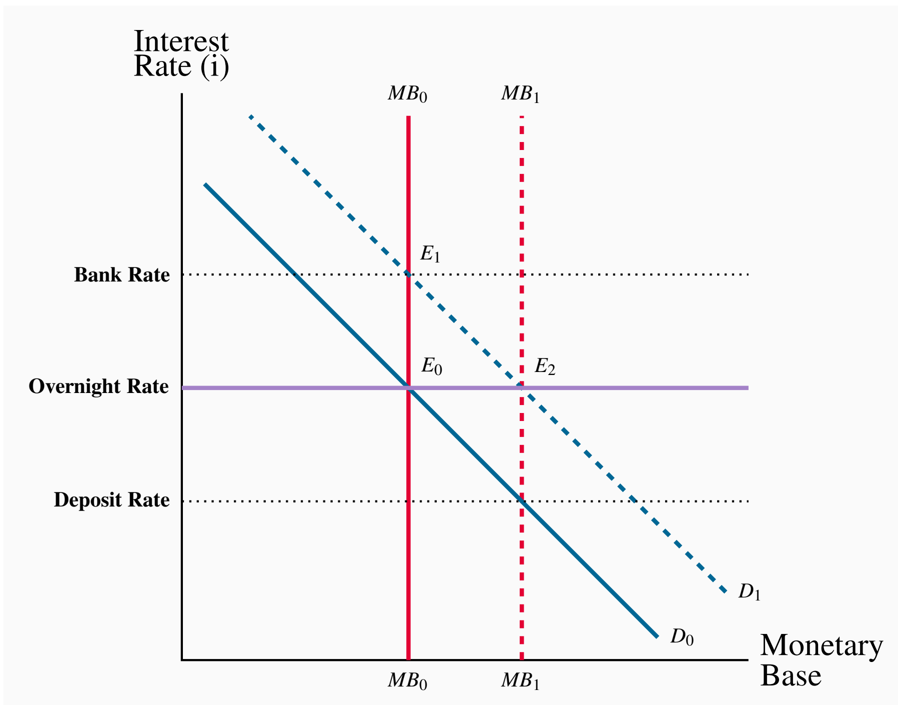
One tool is the **special purchase and resale agreement (SPRA)**. This is a transaction initiated by the Bank of Canada that puts cash into the system on a very short-term basis. It is used to maintain the target overnight rate, more specifically to offset upward pressure on the rate.

In an SPRA, the Bank offers to buy Government of Canada securities from major financial institutions with an agreement to sell them back the next business day, at a predetermined price. The financial market gets an overnight injection of monetary base. The difference between the purchase and resale price determines the overnight interest rate on the transactions. Banks are willing to enter into these agreements with the Bank of Canada because they provide cash for the banks at rates of interest below what they would otherwise have to pay in the overnight market.

> **SPRA**: A Bank of Canada purchase of securities one day combined with an agreed resale of the securities the next day.

Figure 22.4 shows how this works. We start in equilibrium at E_0 with the demand for cash reserves just equal to the supply of monetary base at the overnight rate set by the Bank. Suppose a change in economic and financial circumstances causes a *temporary* increase in the demand for cash and settlement balances. The demand for monetary base shifts to the right to D_1. If the Bank took no action the overnight rate would rise to E_1 at or even above the target band the Bank has set for the overnight rate. To prevent this, the Bank provides an overnight increase in the monetary base by buying securities on the agreement that it will sell them back the next business day. This is an SPRA. It gives a temporary increase in monetary base to MB_1 and reinforces the Bank's overnight rate.

Figure 22.4: Setting the overnight interest rate



In the opposite case of a fall in the demand for monetary base, the Bank makes use of a second tool to reduce the monetary base; namely, a **sale and repurchase agreement (SRA)**. This is a sale of securities to major financial institutions for one day combined with a repurchase the following day. It makes a one day reduction on the monetary base to offset a drop in the demand for cash that has put downward pressure on the overnight rate.

> **SRA**: A Bank of Canada sale of securities one day combined with an agreed repurchase of the securities the next day.

What if the increased demand for monetary base is permanent, not just a one-day event? Chapter 20 showed how the demand for monetary base comes from public demand and from bank demand for cash and reserve balances. Increases in nominal income increase the demand for money, including cash. Suppose the increased demand for monetary base shown in Figure 22.4 is permanent, the result of growth in the economy. Now, if the Bank of Canada wants to keep the overnight rate constant, it must make a permanent increase in the monetary base, not just an overnight increase. It will do this using the open-market operations discussed earlier. An open-market purchase of government securities makes a lasting increase in the Bank of Canada's asset holdings, permanently increasing the monetary base.

While these operations by the Bank of Canada may seem quite complex, they have a simple effect. Like open-market operations, SPRAs and SRAs increase or decrease the monetary base. In this case, however, the buyers and sellers in the market are limited to a few major financial institutions rather than the full market. Furthermore, changes in the monetary base are very short term, just one day. The objective is to set and control short-term interest rates, the *monetary policy instrument* used by the Bank of Canada.

22.4 Monetary policy rules

Most central banks now use interest rates as the main instrument of monetary policy. But how does a central bank decide what rate of interest to set and when to change its settings of the interest rate? What lies behind the announcement of the overnight rate by the Bank of Canada or the setting of the federal funds rate in the United States? How are these interest rate decisions related to economic variables? Professor John Taylor of Stanford University found that most central banks, in fact, adjust interest rates in response to changes in two variables, output and inflation.

This finding was contentious. It implied monetary supply targets no longer played a role in decisions about setting interest rates. Instead, the interest rate target was and is set based on expected inflation and expected output relative to the central bank's inflation target and the economy's potential output.

A central bank that follows a Taylor rule cares about output stability as well as price stability. However, the basic aggregate demand and supply model in Chapter 17 showed that deviations of output from potential output had predictable effects on prices and inflation rates. Booms leading to inflationary gaps push prices up and lead to inflation. Recessionary gaps tend to reduce inflation. Thus, a Taylor rule is also compatible with the interpretation that the central bank cares about prices and inflation, both now and in the future. It is hard to distinguish empirically between these two interpretations of why a Taylor rule is being followed. Indeed, recently inflation rates in most industrial economies have been very low and central banks have come to focus more on output gaps and corresponding unemployment rates in making their interest rate setting decisions.

Nevertheless, until the financial crisis of 2008, Taylor's claim that such a rule effectively described central bank policy had strong empirical support. Most of the leading central banks, including the US Federal Reserve, the Bank of England, and the new European Central Bank used an interest rate target as a policy instrument in pursuit of an inflation control objective. Taylor's insight is so widely used that it is called the "**Taylor rule**". It still provides a useful way to explain monetary policy decisions.

Taylor rule: central bank interest rate settings based on inflation and output targets.

A simple monetary policy rule

It is useful to assume *prices are constant*, so the inflation rate is zero for the first part of the study of policy rules. When prices are constant, monetary policy follows the output part of a policy rule.

In simple terms, the rule is:

The overnight rate that is set (i_{onr}) equals:

> (**i**$_0$) the **neutral rate of interest** consistent with equilibrium at potential output (Y_P) *plus an adjustment to that rate if*:

There is currently an output gap $(Y - Y_P) \neq 0$, which gives the equation:

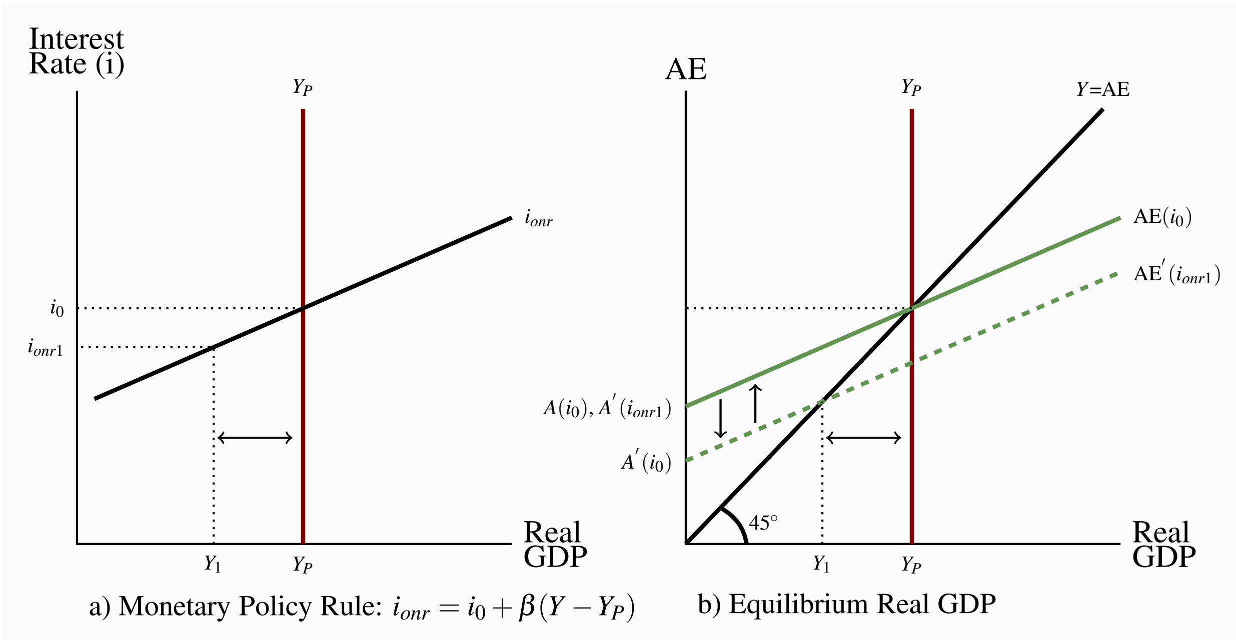
$$i_{onr} = i_0 + \beta(Y - Y_P) \tag{22.1}$$

When real output (Y) is at potential output (Y_P), the output gap is zero and the interest rate is set at $i_{onr} = i_0$. That is the central bank's judgment on the interest rate required to support aggregate demand and equilibrium at potential output. If, according to this rule, output temporarily exceeds Y_P, $[(Y - Y_P)/Y_P] > 0$, the central bank raises interest rates to reduce aggregate expenditure. At levels of output below Y_P, it lowers interest rates. The size of the difference between i_{onr} and i_0 is determined by the size of the parameter β, which reflects the central bank's estimate of the size of the effect of an interest rate change on aggregate expenditure.

> **Neutral rate of interest**: the interest rate that, in the central bank's judgment, is consistent with actual real GDP equal to potential real GDP and inflation at target.

Figure 22.5 shows how this policy rule works. In Panel a), output is measured on the horizontal axis and interest rates on the vertical axis. A vertical line is drawn at the potential output level Y_P. The positively sloped line showing the central bank's reaction to fluctuations in output crosses the Y_P line at the interest rate i_0 with a slope of β. If output were lower than Y_P, at Y_1 for example, the bank would lower interest rates to i_1 to provide some stimulus to aggregate expenditure.

Figure 22.5: Interest rates and output with a simple monetary policy rule



Panel a) The Central bank sets overnight interest rate at the neutral rate i_0 consistent with Y_P. If Y does not equal Y_P the Bank changes the overnight rate to either add or reduce monetary stimulus. Slope of line defines central bank reaction to any output gap, $Y \neq Y_P$.
Panel b) A fall in $A(i)$ causes a recessionary fluctuation in Y to Y_1. Bank reacts, reducing the overnight rate to i_{onr1}. Lower i increases $A(i)$ to restore AE and equilibrium $Y = Y_P$. If decline in A is persistent and not offset by the initial cut in i_{onr} the Bank can cut its overnight rate again.

Alternatively, for an output greater than Y_P, the central bank would raise the overnight interest rates to reduce aggregate expenditure. This simple policy rule describes how the central bank resets its interest rate to achieve the target of equilibrium output at Y_P, and reacts to offset or moderate *fluctuations* about Y_P.

Panel b) shows the effect of the changes in interest rates on equilibrium output. We studied the transmission mechanism from interest rate changes to expenditure changes in Chapter 21. The central bank chooses the overnight interest rate i_{onr} that gives aggregate expenditure AE(i_{onr}) and equilibrium output at Y_P.

If *short-run* economic conditions changed and autonomous expenditure declined, the AE line would shift down to AE$'(i_0)$. Y would fall to Y_1, which is less than Y_P. From the policy rule in Panel a), we see the reaction of the central bank. It lowers the overnight interest rate to i_{onr1}. Lower interest rates work through the cost of and availability of finance, wealth effects, and exchange rate effects to increase expenditure. The AE line shifts back up to AE$'(i_{onr1})$ to restore equilibrium at Y_P. By reacting to temporary changes in the state of the economy, the central bank attempts to stabilize output at Y_P.

The simple policy rule in Figure 22.5 sets an overnight interest rate target i_0 based on an assessment of the *fundamental conditions* in the economy that underlie potential output. It calls for changes in the interest rate in response to temporary fluctuations in those conditions.

A more lasting change in economic conditions would call for a reassessment of the neutral rate of interest (i_0) and a new policy rule. The Bank of Canada's current range for the neutral interest rate (i_0) is 2.5% – 3.5%. If longer term projections for sustained economic conditions changed, the Bank would revise its estimate of the neutral rate. That would change i_0 in the policy rule and shift the policy line in the diagram. Announcements of changes in the overnight rate in Canada or the federal funds rate in the United States usually come with a brief comment on current and expected economic conditions to give financial markets information on likely changes in interest rates going forward.

A large external shock that disrupted economic conditions, such as COVID-19 and policy responses to shutdown and distancing, would displace the idea of a neutral rate other than the 'effective lower bound' defined below.

Policy rules and inflation

The assumption that prices are constant ignores the importance of inflation, making a policy rule based on the output gap alone too simple. Central banks today, like the Bank of Canada, conduct their monetary policy by setting *inflation targets*. This does not mean that they ignore the level of output in the economy. Instead, inflationary and recessionary gaps, and changes in unemployment rates, are seen as important predictors of future inflation. The simple monetary policy rule of Equation 22.1 is easily extended to recognize this current approach to monetary policy.

The central bank's setting of its interest rate instrument could then be:

1. (i_0) the neutral interest rate consistent with equilibrium at potential output (Y_P) plus an adjustment to that rate if:
2. (π) the current inflation rate differs from π^*, the Bank's inflation rate target plus a second adjustment if:
3. $[(Y - Y_P)/Y_P]$ the output gap does not equal zero.

This gives a basic equation that describes the central bank's setting of the overnight rate in reaction to economic conditions.

$$i_{onr} = i_0 + \alpha(\pi - \pi^*) + \beta[(Y - Y_P)/Y_P] \tag{22.2}$$

As before, the central bank estimates a neutral interest rate i_0. This is the nominal interest rate the bank thinks is consistent with output at potential output and inflation at the target rate π^* under current conditions.

The Bank of Canada's current inflation target, for example, is $\pi^* = 2.0$ percent, the midpoint of a 1 percent to 3 percent range. If inflation rises above the target π^*, the central bank raises the overnight interest rate. The parameter α in the equation tells us by how much the overnight interest rate would be changed, either up or down in response to an inflation rate different from the bank's target.

Expenditure decisions depend on the interest rate. To stick to its inflation target, the bank must change the interest rate by changing the overnight interest rate by more than any change in inflation. This requires the parameter $\alpha > 1$. A rise in inflation is then met by a rise in interest rates that is large enough to reduce expenditure and inflationary pressure.

By this rule, the central bank also reacts to any departure of output from potential output, as it did in our earlier simple rule, Equation 22.1. The parameter β measures how much the central bank would raise the overnight interest rate in response to an inflationary gap, or lower it in response to a recessionary gap. Output stabilization requires that interest rates be raised in the case of an inflationary gap or lowered in the case of a recessionary gap thus β is greater than zero ($\beta > 0$).

Changing interest rates to offset an output gap is intended to stabilize output, but it will also work to offset any changes in the *future* inflation rate that would be caused by a persistent output gap. The size of the central bank's reactions, as measured by the parameters α and β are indications of the relative importance it attaches to inflation control and output stabilization.

Any change in economic conditions that the central bank thinks is *going to last for some time* results in a new estimate of the neutral rate i_0. In that case the monetary policy line in a diagram would shift up or down. Interest rates would then be higher or lower for all inflation rates and output gaps, depending on the change in i_0. The central bank would explain this change when announcing the change in the setting of its policy instrument, the overnight rate in Canada or the federal funds rate in the United States. Bank of Canada press releases explain overnight rate settings. See for example:

http://www.bankofcanada.ca/2017/04/fad-press-release-2017-04-12/

This approach to monetary policy has similarities to our earlier discussion of fiscal policy. In that case, we distinguished between automatic and discretionary policy. In the case of monetary policy, the discretionary component is the setting of its estimated *neutral* interest rate based on an evaluation of longer-term economic conditions relative to the target inflation rate. It *positions* the monetary policy line in a diagram in much the same way as the structural budget balance *positions* the government's *BB* line in Chapter 19. Short-term fluctuations in economic conditions result in short-term variations in the overnight rate—movements along the monetary policy line. This is similar to the automatic stabilization that comes from movements along the government's *BB* line as a result of fluctuations in output and income.

There is, however, an important difference between monetary and fiscal policy. Monetary policy *that uses the interest rate as the policy instrument provides strong automatic stabilization in response to money and financial market disturbances. Automatic stabilization in fiscal policy reduces the effects of fluctuations in autonomous expenditures.*

The effective lower bound (ELB)

The financial crisis and recession of 2008-09 led to new and more intense monetary policy actions by central banks. Most continued with cuts to basic policy rates as their first response. The Federal Reserve in the United States lowered its federal funds rate, in steps, to a range of zero to 0.25 percent. The Bank of Canada followed, lowering its overnight rate setting to 0.5 percent by early March 2009. But these lower rates were not sufficient to stimulate borrowing and expenditure. Banks and other lenders were concerned by the increased risks of losses on their current lending and the risks involved in new lending. They had suffered losses on previous large denomination fixed term and mortgage lending. Bankruptcies were rising across many business and consumer loan markets.

In response to the COVID-19 pandemic in early 2020, the Bank of Canada cut its target for the overnight interest rate from 1.75 percent to 0.25 percent. That is as low as the overnight rate can go, namely the effective lower bound.

With their policy interest rates cut to near zero, central banks hit the **effective lower bound (ELB)**. Nominal interest rates could not be reduced any further. Central banks needed additional policy tools to meet deep concerns about risk and liquidity in financial markets. Increased demands for liquidity raised desired reserve and currency holdings and lowered money supply multipliers in many countries, and restricted access to bank credit.

> **Effective lower bound (ELB)**: A Bank's policy interest rate cannot be set below a small positive number.

Two previously used techniques were introduced. The first was increased "**moral suasion**," an increase in communications with financial market participants to emphasize the central bank's longer-term support for markets and its actions to promote stability. More directly the banks were urged to maintain their lending operations.

> **Moral suasion**: a central bank persuades and encourages banks to follow its policy initiatives and guidance.

The second was "**quantitative easing**," and in the case of the US, an even more extensive "**credit easing**." "Quantitative easing" is the large scale purchase of government securities on the open market. It expands the central bank's balance sheet and the size of the monetary base. A version of this policy action was used in Japan earlier in the decade after the Bank of Japan had lowered its borrowing rate to zero and wanted to provide further economic stimulus. The Bank of Canada began using quantitative easing in April 2020, to provide liquidity and financial support for the economy in addition to having cut its overnight rate to the effective lower bound. The increase in the Bank's assets and liabilities, shown in Application Box 22.1, illustrate the scale of the resulting increase in the monetary base.

> **Quantitative easing**: a large scale purchase of government securities to increase the monetary base.

Credit easing is measured by the expanded variety of loans and securities the central bank willingly holds on its balance sheet. These come from purchases of private sector assets in certain troubled credit markets. The mortgage market for example was in trouble as a result of falling real estate prices and mortgage defaults. Cash is put directly into specific markets rather than letting it feed it through commercial banks' lending and loan portfolio decisions.

> **Credit easing**: the management of the central bank's assets designed to support lending in specific financial markets.

Monetary policy practice continues to evolve. In the last few years, with policy interest rates at or near the effective lower bound and persistent weakness in economic growth and employment, major central banks have relied increasingly on "**forward guidance**" to support their economies. Forward guidance is contained in the explanation a Bank gives in its formal announcement of setting of the policy interest rate. If for example the Bank's opinion that economic growth and inflation will be slow and weak, the Bank suggests that interest rates will be unchanged for some time into the future. Alternatively a prediction of a revival in growth and inflationary pressure may lead the Bank to predict increases in policy interest rates in the near future.

> **Forward guidance**: information on the timing of future changes in the central bank's interest rate setting.

This forward guidance is intended to help firms and households make expenditure decisions that require debt financing. In some cases it is based on an explicit economic criterion.

For example, the US and the UK introduced unemployment rate thresholds for changes in the monetary policy rate settings. This is in essence a variety of forward guidance. In the UK, in August 2013 the Bank of England, under the heading "Forward Guidance" announced:

> *In particular, the MPC [Monetary Policy Committee] intends not to raise Bank Rate from its current level of 0.5% at least until the Labour Force Survey headline measure of the unemployment rate has fallen to a threshold of 7%, subject to the conditions below.*

Source:

https://www.bankofengland.co.uk/minutes/2013/monetary-policy-committee-august-2013

In effect, putting an unemployment rate target into the interest rate setting rule is a variation on the rule specified by Equation 22.2. It adds to or replaces the output gap target with an unemployment rate target.

22.5 Monetary policy indicators

Policy rules describe how a central bank, like the Bank of Canada, would use interest rates to stabilize output, prices, and inflation in the economy. To see how the Bank's actions affect economic activity and inflation, we need some indicators of the expansionary or restrictive stance of monetary policy. These **monetary policy indicators** will allow us to go beyond the central bank's descriptions of its policy and observe the effects of its policy actions on monetary conditions in the economy.

> **Monetary policy indicators**: variables that provide information about the stimulus or restraint coming from the central bank's policy.

Our earlier discussion of the monetary transmission mechanism suggests two monetary policy indicators, namely, *interest rates* and *exchange rates*. The central bank sets nominal interest rates, which have important effects on asset prices, cash flows, and expenditures. Interest rates are also important to expenditure decisions. Changes in nominal interest rates over time will show how monetary policy has been implemented.

The foreign exchange rate, as it affects net exports, also provides an indicator of policy stance. Because exchange rates change in part as a result of interest rate differences between countries, changes in the exchange rate provide an indicator of the thrust of domestic monetary policy relative to foreign monetary policy. Although in Canada it is important to recognize that commodity prices also have strong exchange rate effects as illustrated by the depreciation of the Canadian dollar when oil and commodity prices fall.

The monetary transmission mechanism works through both interest rates and exchange rates. In setting its interest rates, a central bank in a small open economy needs to consider recent changes in the exchange rate. If economic conditions, or policies in other countries, have caused changes in the foreign exchange rate, those changes will affect expenditures and output in Canada.

The depreciation of the Canadian dollar in 2015 is an important example. The corresponding appreciation of the US dollar raised import prices and increased the profitability of exports. Even without monetary policy action, expenditure and output in Canada would rise. The Bank of Canada had to make a decision. Was the setting of its operating range for the overnight rate still consistent with its inflation target once the exchange rate had risen? Should the Bank count on the stimulus from the exchange rate to offset the effects of lower energy and commodity prices or cut its interest rate to provide some further stimulus? If it were to respond, by how much should it lower interest rates? The Bank of Canada lowered interest rates. In this case the combined effect of interest rates and exchange rates was increased stimulus. However, the Bank could have decided that the exchange rate alone would suffice or even that the exchange rate depreciation gave too much stimulus. Clearly both are very important for designing and judging and monetary policy.

While interest rates and exchange rates provide important indicators of monetary policy, many economists and the Bank also regard the money supply or the rate of growth of the money supply as a policy indicator. Some suggest a monetary policy rule for money supply, which uses money supply as the central bank's policy instrument. The demand for nominal money balances depends on nominal income. Taking this into account, the difference between the rate of growth of the money supply measure, M1B, and the rate of growth of nominal GDP provides an indicator of the stance of monetary policy. M1B growth that exceeds growth in nominal GDP provides easier financial market conditions and suggests an expansionary policy stance.

The growth rates in the money aggregates M1B+ and real M2+, adjusted for inflation, provide alternative indicators of the effect of monetary policy. In the current policy context, the Bank of Canada sets the interest rates and the growth rates of money supply reflect the demand for money balances at those interest rates. Empirical research at the Bank and by other monetary economists has found that the growth in real M1B+ is a useful indicator of future growth in real GDP. Growth in real M2+ also provides a leading indicator of inflation. From these findings, an observed increase in the growth rates of these money aggregates indicates that the Bank's current policy is adding to aggregate demand.

Thus we have a basic set of monetary policy indicators: interest rates, exchange rates, and the growth rate in nominal and real measure of money supply. They come from our understanding of the way changes in monetary variables may affect expenditures, incomes, and prices and from our discussion of how monetary policy is designed and implemented. The Bank of Canada provides a more extensive list of key monetary policy variables at:

http://www.bankofcanada.ca/rates/indicators/key-variables/

Next

This chapter completes the development and explanation of the basic expenditure and monetary structure of the economy under the assumption that the general price level is constant. Chapter 23 introduces a basic modern AD/AS model that explains short-run fluctuations, output, and inflation. That model is built on modern monetary policy that sets interest rates to achieve inflation control. It is used to illustrate and evaluate recent Canadian fiscal and monetary policies.

Key Concepts

Central banks operate to influence the behaviour of other banks and intermediaries in the financial system.

A central bank conducts **monetary policy** through its control of the monetary base and interest rates. It is also banker to the government and to the commercial banks.

The **Bank of Canada** is Canada's central bank. It is the source of the monetary base. It sets short-term interest rates, acts as banker to the commercial banks and the federal government, and is the lender of last resort to the banks.

Monetary policy in Canada is the responsibility of the Bank of Canada. The Bank uses its control of the monetary base and interest rates to promote economic stability at potential output and a low stable inflation rate.

Central banks have three main operating techniques: **reserve requirements imposed on commercial banks**, **open-market operations**, and **bank rate** setting. These techniques are used to manage the monetary base, the money multiplier, and interest rates.

Central banks can implement **monetary policy through the monetary base and money supply control** or through interest rate control, but cannot do both simultaneously.

In practice, the Bank cannot control money supply exactly. Thus, for most central banks, a short-term **interest rate is the instrument of monetary policy**.

The Bank of Canada uses the **overnight interest rate as its policy instrument**, and an inflation rate of **1 percent to 3 percent as its policy target**.

The Bank of Canada uses **SPRAs** and **SRAs** to intervene in the market for overnight funds and to reinforce its setting of the overnight interest rate.

A **monetary policy rule** such as a rule for setting the interest rate provides a useful description of the way the central bank sets and adjusts its interest rate policy instrument.

The **neutral rate of interest** is the interest rate that, in the central bank's judgment, is consistent with actual real GDP equal to potential real GDP and inflation at target.

Changes in the central bank's policy instrument change nominal and real interest rates and change aggregate demand through the **transmission mechanism**, which includes wealth effects, cost of financing effects, and exchange rate effects on the components of aggregate expenditure.

Quantitative easing is the use of central bank purchases of securities with the aim of increasing the monetary base to meet unusually high demands for liquid cash balances in times of financial and economic crisis.

Credit easing is the increase in specific kinds of central bank asset holdings (for example, commercial paper) designed to provide liquidity and support lending in specific markets facing shortages of funds.

Forward guidance: information on the timing of future changes in the central bank's interest rate setting.

Real and nominal interest rates, exchange rates and rates of growth of money aggregates relative to national income can be used as **monetary policy indicators**.

Exercises for Chapter 22

Exercise 22.1 Explain carefully why a central bank does not operate to make a profit but a commercial bank does. What is the central bank's operating objective? What unique power does a central bank have that allows it to pursue its operating objective?

Exercise 22.2 Explain carefully why a central bank's power to conduct monetary policy is based on its unique position as supplier of the monetary base.

Exercise 22.3 Why would a change in the monetary base ΔMB cause a change in the money supply?

(a) Suppose a central bank buys \$10 million on the open market. What effect does this have on the monetary base and the reserve position of the commercial banks?

(b) If the banks hold reserves equal to 2.5 percent of their deposit liabilities, and the public holds a constant amount of cash, calculate the effect of this open-market transaction on:
   i. The money supply.
   ii. The banks' reserve balances.

Exercise 22.4 Suppose the central bank decides to use its power to set interest rates. Use a money market diagram to show and explain what happens to the real money supply if real output increases $(\Delta Y > 0)$ and the central bank maintains a constant interest rate.

Exercise 22.5 In terms of a monetary policy rule

(a) What is the Bank of Canada's monetary policy target?

(b) What monetary policy instrument does the Bank use to pursue this target?

(c) What do the Bank's procedures for implementing policy mean for its control over money supply?

Exercise 22.6 Use a diagram to show circumstances in the market for overnight funds that might lead the Bank of Canada to make an SRA. Why would the Bank use an SRA in this case rather than an open market operation?

Exercise 22.7 Suppose a central bank decides to conduct monetary policy according to a rule for interest rates.

(a) How does it choose the basic setting for the interest rate within the rule?

(b) How would it respond to an output gap $[(Y - Y_P)/Y_P] > 0$?

(c) How would the bank react to an inflation rate higher than its target inflation rate?

(d) Why would the bank decide to change the basic setting of its interest rate?

Part Nine

Real GDP, Business Cycles, Policy and Growth

This Part integrates the material of the preceding chapters. Chapter 23 builds a 'modern' aggregate demand-aggregate supply model to explain the *inflation rate*, and real GDP and stabilization policy. The current approaches in inflation control, in which monetary policy sets the interest rate, are discussed. Chapter 24 introduces international aspects of macroeconomics and the importance of foreign exchange rates to the design and effectiveness of monetary and fiscal policy, using the model from Chapter 23. Chapter 25 introduces theories of economic growth and the importance of technology and productivity growth for standards of living.

Chapter 23

Inflation, real GDP, monetary policy & fiscal policy

In this chapter we will explore:

Chapter 16 identified the inflation rate, the unemployment rate and the GDP growth rate as key indicators of economic performance. This chapter combines the concepts of earlier chapters into a full basic macroeconomic model of the economy for the study of causes of change in the first two of these indicators. It is representative of the models currently used by central banks and policy analysts to evaluate current economic performance and policy, forecast future performance and prescribe policy action. Economic growth and growth rates are covered in Chapter 25.

Normally, in the absence of random shocks, underlying short-run conditions on the supply side of the economy are the same in this chapter as in earlier chapters. Labour force and technology are fixed by assumption, fixing potential output at Y_P. Instead of a fixed short-term equilibrium price level, the modern model assumes equilibrium rates of change in money wage rates, and inflation rates are stable in the short run and adjust slowly to output gaps.

On the demand side, monetary policy is aimed at inflation control using the short-term interest rate as the instrument as explained in Chapter 22. It reflects current Bank of Canada monetary policy as explained by the Bank at: `www.bankofcanada.ca/core-functions/monetary-policy/`. In normal conditions this allows fiscal policy to focus on managing the government's budget balance as it affects output, employment and the public debt ratio.

The model provides explanations of business cycle fluctuations in output, employment and inflation rates. It provides a basis for evaluating recent changes in economic conditions and monetary and fiscal policy actions. It can also be used to illustrate the changes in equilibrium real GDP and the inflation rate caused by policies to control the spread of the COVID-19 pandemic.

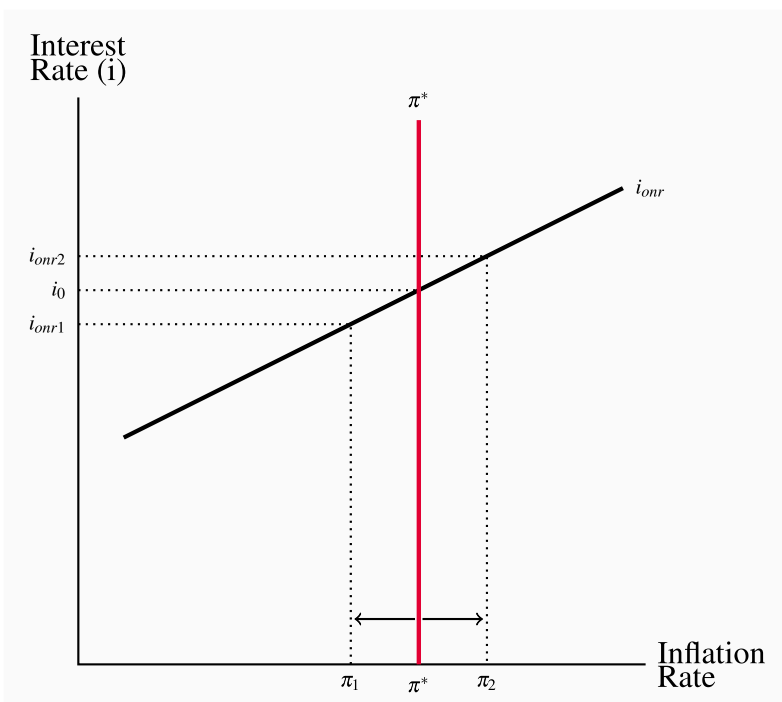
23.1 Inflation and aggregate demand

Monetary policy establishes the link between the inflation rate and aggregate expenditure that determines the *slope of the AD curve*. Central banks set *interest rates* to control the *inflation rate* based on an inflation rate target. A monetary policy that reacts to changes in the inflation rate by changing the interest rate causes changes in expenditures. The link between the inflation rate and aggregate expenditure still comes through interest rates and the monetary transmission mechanism, but it is the central bank's decision to change its policy interest rate that provides the impulse.

Changes in autonomous expenditure that are independent of interest rates are still an important part of aggregate demand. They cause *shifts in the AD* curve based on the multiplier as explained in earlier chapters. Changes in government fiscal policy and in the central bank's estimate of the neutral interest rate also shift AD. It is the monetary policy response to *transitory variations* in the inflation rate that give the AD curve its slope.

Figure 23.1 illustrates this monetary policy. Assume the central bank has an inflation control target such as π^*. Based on its evaluation of current and near future aggregate expenditure and supply conditions the Bank sets the interest rate i_0 to get its target inflation rate. If the actual inflation rate differs from π^* as a result of changes in economic conditions that the Bank views as temporary, rather than fundamental, it reacts with temporary changes in its interest rate within a narrow range, allowing some variation around i_0 such as i_{onr1} to i_{onr2}. The upward sloping line in the diagram shows this 'reaction function'.

Figure 23.1: Monetary policy sets Overnight Interest Rate



A monetary policy rule: $i_{onr} = i_0 + \beta(\pi - \pi^*)$.

This is a simple policy rule you will recall from Chapter 22. An equation that describes this policy rule would be:

$$i_{onr} = i_0 + \beta(\pi - \pi^*) \tag{23.1}$$

In this rule i_0 is the neutral interest rate. The Bank's target inflation rate is π^*. An economy working at potential output would have a constant inflation rate π^* at i_0.

The parameter β defines the size of the change the Bank would make to its interest rate in response to differences between the observed inflation rate and the target rate. β is the slope of the overnight interest rate line (i) in the diagram. If the inflation rate were to rise to π_2 in Figure 23.1 the Bank would raise its overnight interest rate above i_0 to i_{onr2} to defend its inflation rate target. The larger is β, the stronger the Bank's reaction to short-term variation in the inflation rate and the steeper the interest rate line.

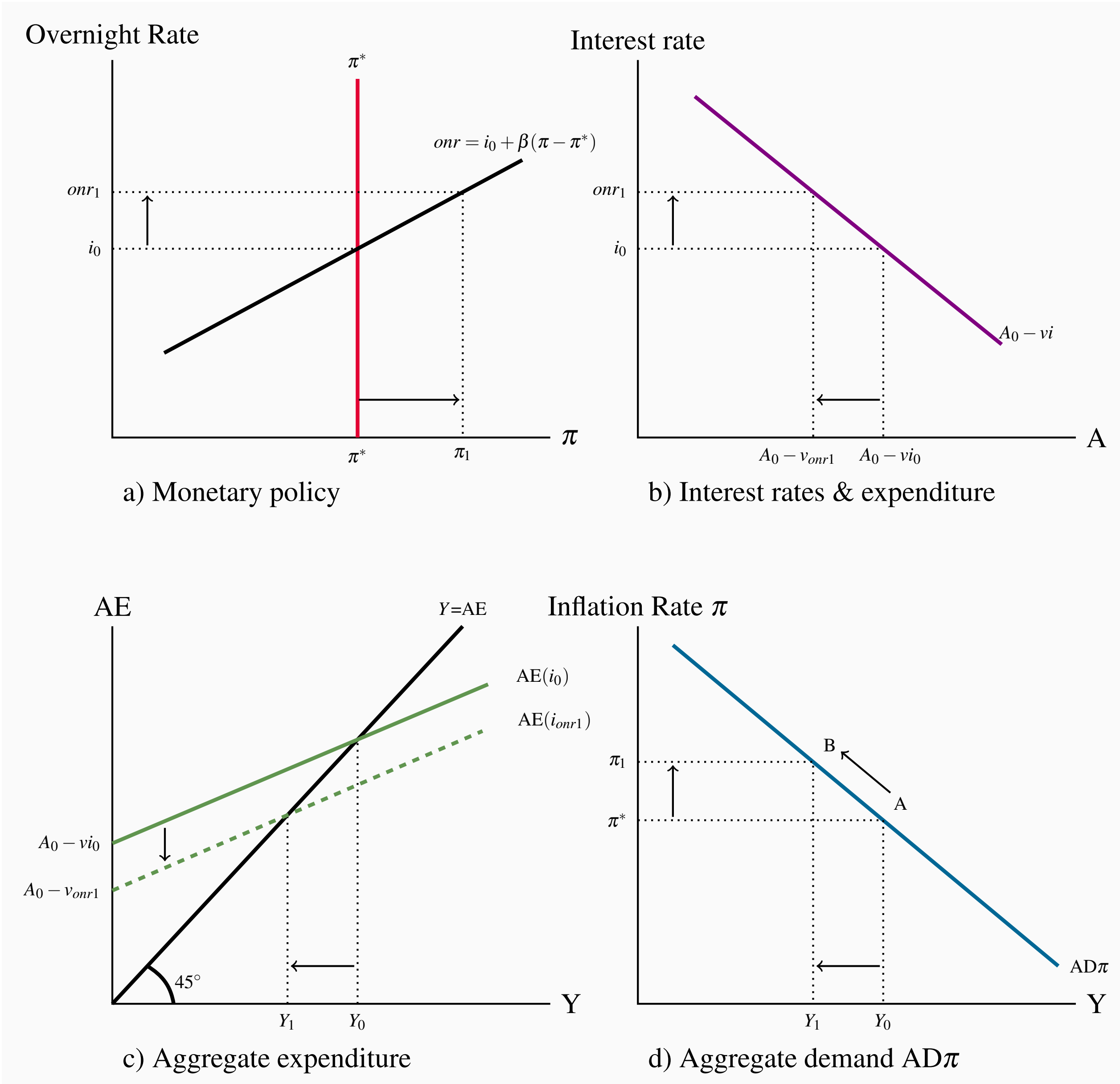
Alternatively if the Bank does not react to temporary departures from π^*, that is if $\beta = 0$, the interest rate line is horizontal. Changes in the Bank's official setting of the neutral overnight rate would then shift the interest rate line up or down as the case may be.

This approach to monetary policy and the inflation rate and aggregate demand is based on two familiar relationships:

1. The monetary transmission mechanism as in Chapter 21, and
2. The expenditure multiplier from Chapters 18 & 19.

Figure 23.2 shows the derivation of the aggregate demand curve labelled ADπ. This is to differentiate it from the AD curve in Figure 21.5. In that example and examples in earlier chapters in which the equilibrium price level was constant, the inflation rate was zero, and a change in the money supply *shifted the AD curve*.

Figure 23.2: The ADπ function



Start in equilibrium with $\pi = \pi^*$ and interest rate $i = i_0$ in a), and π^* and Y_0 at A in d). Suppose the inflation rate rises to π_1. The central bank then raises its overnight interest rate to i_1. The higher interest rate reduces expenditure in b), lowering AE in c) and equilibrium real GDP to Y_1. The new equilibrium combination $\pi_1 Y_1$ is point B on the ADπ curve.

Figure 23.2 illustrates the relationship between the inflation rate, monetary policy and the slope of the ADπ curve. In this example the central bank *does* react to temporary variations in inflation, π, ($\beta > 1$). Monetary policy as described by a policy rule or reaction function like Equation 23.1 is in Panel a) of the diagram. Interest rates set by policy in Panel a) determine autonomous expenditures in Panel b) giving the monetary transmission mechanism. Autonomous expenditures determine the position of the AE line in Panel c) and work through the multiplier to determine equilibrium real GDP. The combination of the inflation rate π^* and equilibrium real GDP, Y_0, give one point (π^*, Y_0) on the ADπ curve plotted in Panel d) and labeled ADπ.

The central bank's reactions to changes in the inflation rate, other things constant, move the economy *along* the ADπ function. In Panel a) the Bank reacts to a temporary or *transitory* rise in the inflation rate to π_1 by allowing its policy interest rate to rise to i_{onr1}. This reduces autonomous expenditure in Panel b) to $A_0 - vi_{onr1}$ and lowers the AE curve in Panel c) to AE(i_{onr1}). The multiplier then lowers equilibrium real GDP to Y_1 and plotting (π_1, Y_1) in Panel d) gives a second point on the ADπ curve labeled B.

Changes in autonomous expenditures (ΔA) not caused by changes in interest rates, like changes in autonomous consumption, investment, exports, or government expenditure would shift the ADπ function by amounts driven by the multiplier. Similarly, a *change in monetary policy made by changing the assumed neutral rate* i_0, in response to a change in economic fundamentals would *shift the ADπ function*.

The recession of 2009, for example, brought a sharp drop in Canadian domestic investment and exports and a shift in ADπ to the left. The Bank of Canada lowered its setting for the overnight interest rate in steps from 4.75 percent to 0.25 percent, as illustrated in Figure 22.2, to offset some of this drop in ADπ. However the monetary stimulus to ADπ through lower interest rates was not enough by itself to avoid a recession. Fiscal stimulus was also needed.

The government policies used in early 2020 to control the spread of COVID-19 disrupted both aggregate demand and aggregate supply. Once again the Bank of Canada cut its overnight rate to 0.25 percent, but any link between interest rates to expenditures was lost to large business closings and employment cuts. Quantitative easing by open market operations were used to help in the financing of large fiscal stimulus programs.

23.2 Aggregate supply

The supply side of the economy explains output, *inflation*, and the economy's adjustment to equilibrium at Y_P, potential output. If the economy is operating with an output gap, changes in the *rate of increase in wage rates* and other factor prices may push the economy toward a long-run equilibrium at potential output and a constant rate of inflation. But this internal adjustment process is slow at best.

No matter what time frame we use, the economy's output depends on:

1. The state of technology;
2. The quantities of factor inputs to production (labour, capital, land, energy, and entrepreneurship); and
3. The efficiency with which resources and technology are used.

A simple **production function** defines the relationship between outputs and labour and capital inputs to production as follows:

$$Y = A \times F(N,K) \tag{23.2}$$

In this equation, Y is real GDP, A is the state of technology, and N and K are, in simplest terms, inputs of labour, measured in terms of hours worked, and capital, measured in terms of physical plant and equipment, respectively, used in the production process.

> **Production function**: outputs determined by technology and inputs of labour and capital.

The notation $F(...)$ tells us that the size of output as measured by real GDP depends on the amount of labour and capital used in the production process. More labour and more capital used means more output. An improvement in technology would make A larger, and increase the output produced by any given amount of labour and capital employed. This would be an increase in **productivity**, an increase in output per unit of input.

> **Productivity**: output per unit of input.

Long-run aggregate supply (Y_P)

Potential output (Y_P) is determined by the current state of technology, A_0, the current stock of capital, K_0, and the *equilibrium* level of employment, N_F. In terms of the simple production function, this means:

$$Y_P = A_0 \times F(N_F, K_0)$$

> **Potential output** (Y_P): the real GDP the economy can produce on a sustained basis with current labour capital and technology without generating inflationary pressure on prices.

The short-run fluctuations in output studied in earlier chapters are linked to differences between actual labour input N and the "full employment" labour input N_F. Unemployment rates fluctuate as a result of these changes in actual output and employment relative to potential output.

Short-run aggregate supply

Short-run aggregate supply in the 'modern' model defines the relationship between output and the inflation rate, when capital stock, technology and the *rate of growth in money wages* are fixed. The basic argument is that monopolistically competitive firms set prices at a mark-up over marginal costs of production. The rates of increase in marginal costs of production depend in turn on the rates of increase in money wage rates, prices of material inputs, and productivity.

Prices are sticky in the very short run. Changing prices is costly and competition among producers means relative prices are important to market position. Furthermore, stable prices may help to build customer loyalty. As a result, price changes only follow changes in costs that last beyond the current short term. A horizontal short-run aggregate supply curve as in Figure 23.3 captures this output-inflation relationship.

Persistently strong aggregate demand would increase employment, output and capacity utilization, and lower productivity. Employment would rise and the unemployment rate would fall below the **NAIRU**, increasing the growth in money wage rates. This upward pressure on current and future marginal costs would compress producers' markups and profitability. Alternatively, persistently weak aggregate demand would lower rates of increase in money wage rates, lower material costs, lower marginal costs, and increase mark-ups. Producers wouldn't adjust price setting practices immediately in either case. But they would react if their mark-ups are persistently pushed away from what they see as the profit maximizing level.

> **NAIRU**: the 'non-accelerating inflation rate of unemployment' that corresponds to N_F at Y_P.

These short-run aggregate conditions are described in a simple ASπ function that explains the current inflation rate as:

1. (π_{t-1}) The rate of inflation last period or last year.
2. $[(Y-Y_P)/Y_P]_{(t-1)}$ The output gap last period or last year.
3. (σ) A term that captures the effects of disturbances such as commodity price shocks or extreme climate conditions or political events that would shift supply conditions.

These factors are combined in an equation along with the parameter γ which measures the size of the effect an output gap has on the inflation rate:

$$\pi_t = \pi_{-1} + \gamma[(Y-Y_P)/Y_P]_{-1} + \sigma \tag{23.3}$$

Equation 23.3 gives a horizontal short-run ASπ curve that is sometimes called an 'inflation adjustment' function. The current inflation rate (π_t) is 'last year's' inflation rate π_{-1} adjusted up or down in reaction to 'last year's' output gap. For example, suppose:

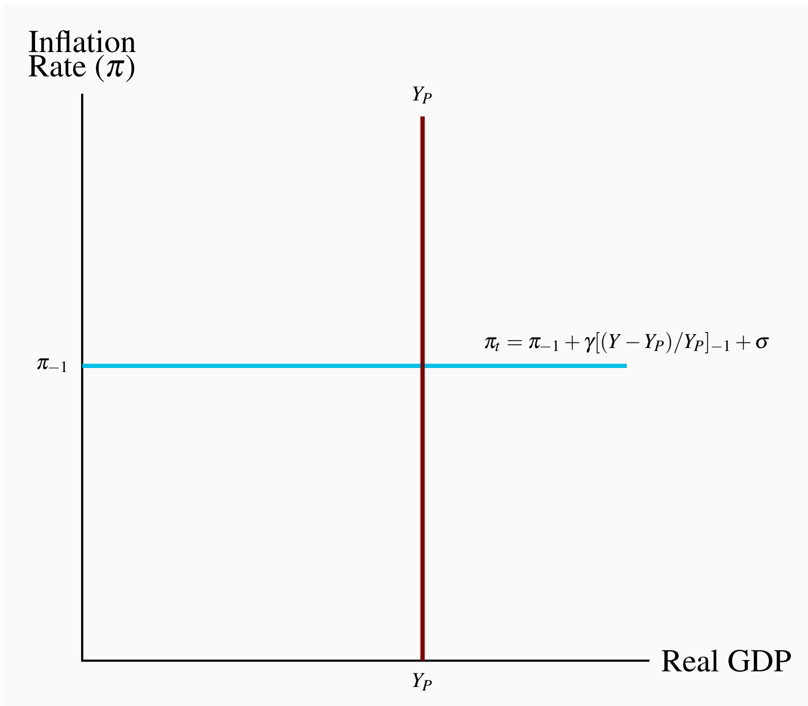
- Last year's inflation rate $\pi_{t-1} = 2.0\%$
- Last year's output gap $[(Y-Y_P)/Y_P]_{(t-1)} = -1.5\%$
- Parameter $\gamma = 0.5$ and $\sigma = 0$

This year's inflation rate π_t would be: $2.0 + (0.5x - 1.5) + 0 = 2.0 - 0.25 = 1.75\%$.

The negative or recessionary output gap in this example has lowered the inflation rate this year to 1.75%.

In other words the inflation rate changes with a lag in response to the effect of an output gap on costs and prices. If there is no output gap, the inflation rate would be the same from year to year. As an alternative, ASπ could be specified as an upward sloping function by dropping the lag on the output gap, but that would remove some of the short-run stickiness observed in wage rates and prices. Figure 23.3 illustrates potential output (Y_P) and short-run aggregate supply ASπ using equation 23.3.

Figure 23.3: Potential output and short-run AS

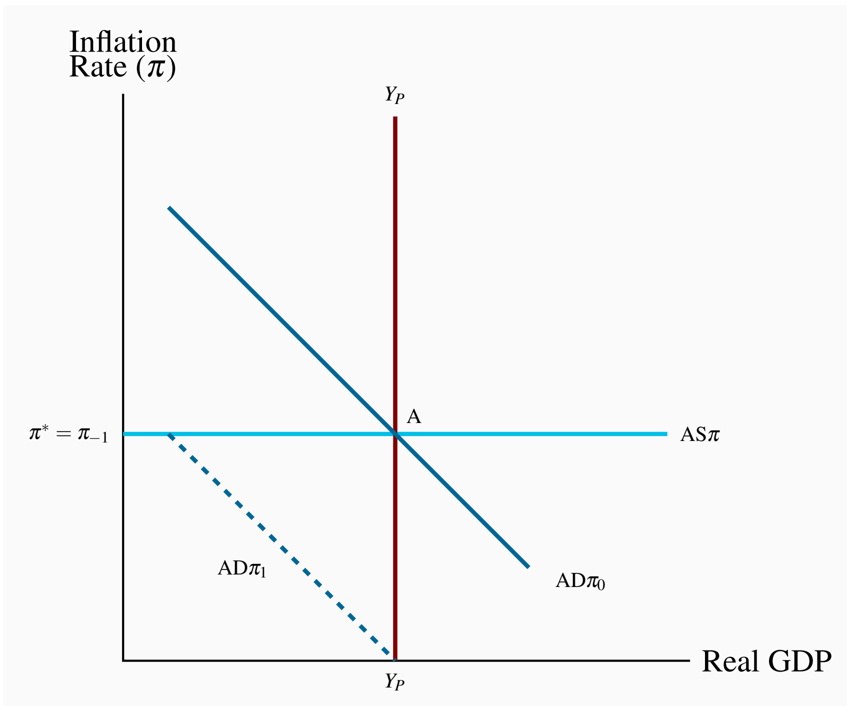


In the short run the inflation rate is constant as long as the economy is at Y_P. If the economy is not at Y_P the output gap will gradually change the π rate, shifting the line up or down in the next time period.

23.3 The equilibrium inflation rate

Figure 23.4 shows the aggregate demand curve ADπ_0 and the vertical long-run aggregate supply curve Y_P. Output is at potential output, and the *inflation rate*, π^*, is determined by aggregate demand. At point A there is equilibrium in all markets: for output, money, and labour.

Figure 23.4: The equilibrium inflation rate



The equilibrium inflation rate is determined by AD_{π_0} and Y_P, as at point A in the diagram. This is also the short-run equilibrium inflation rate. If π^* is the central bank's monetary policy target, the bank sets its interest rate to get AD_{π_0}.

The position of aggregate demand curve ADπ_0, is determined by autonomous expenditure, fiscal policy, and the monetary policy of the central bank.

If monetary policy were directed to a constant equilibrium price level, as in earlier chapters, the equilibrium inflation rate is zero. This is shown in Figure 23.4 by the aggregate demand curve ADπ_1. Then equilibrium is at output Y_P and the inflation rate $\pi = 0$. This ADπ_1 curve is not extended below a zero inflation rate. The discussion of deflation, negative inflation rates, on output and policy are important issues covered in Section 23.6.

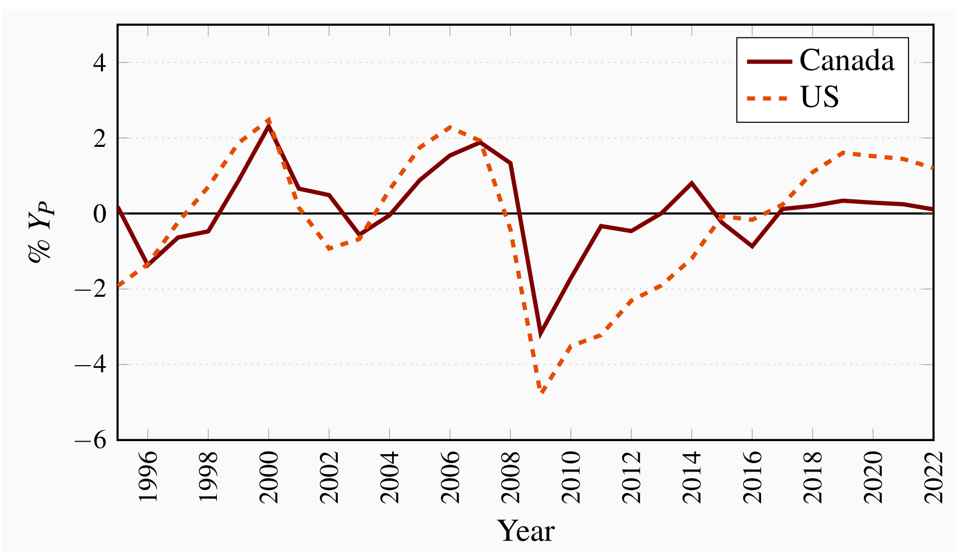
In practice, central banks do not set zero inflation targets for monetary policy. Some, including the Bank of Canada, set explicit low inflation rate targets. Others work to implicit inflation rate targets that are also positive and low. There has been and still is a lot of discussion among economists about the appropriate level of the inflation rate target. The Bank of Canada's 2 percent target represents a current consensus on the issue.

The ADπ_0 curve in Figure 23.4 is based on a monetary policy inflation target of π^*. In setting its inflation target, the central bank recognizes that *money is neutral* when wages and prices are flexible and there is no money illusion. This means that the central bank cannot influence potential output, but it can determine the equilibrium inflation rate. It sets the interest rate and accepts growth in the money supply consistent with its inflation target. A rate of inflation, $[(P_t - P_{t-1})/P_{t-1}]$, greater than zero means the rate of growth of the money supply, $[(M_t - M_{t-1})/M_{t-1}]$, is greater than zero. This puts the ADπ curve at ADπ_0, and keeps it there as inflation raises the price level at the target rate of inflation.

23.4 Adjustments to output gaps

Now consider how demand or supply shocks that move the economy away from potential output trigger an *adjustment process*. In combining the ADπ and ASπ curves, it is assumed the market for goods and services clears even in the short run, but the labour market takes longer to adjust. The inflation rate changes over time as wage rate growth adjusts to the unemployment rate. If this process goes smoothly and other conditions are tranquil, it may eventually restore full employment at potential output. But in the meantime output gaps can be quite persistent.

Figure 23.5: Reported Output Gaps for Canada and the US: 1995–2018 with estimates for 2019–2022



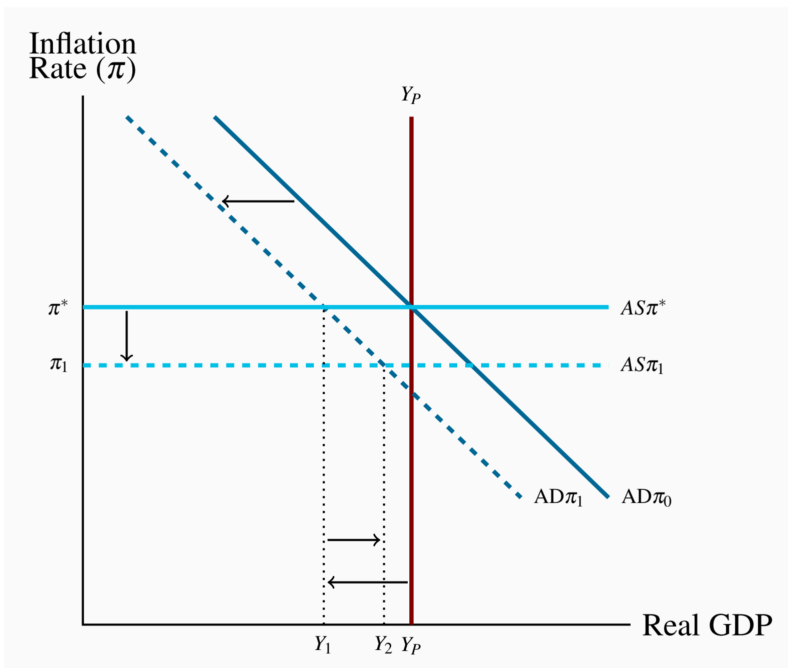
Source: IMF World Economic Outlook Database, October 2018.

Figure 23.5 provides some empirical evidence. For Canada, the recessionary gap of the last half of the 1990s lasted for about five years. The small inflationary gaps that followed persisted for seven years, and were followed by large recessionary gaps in the financial crisis of 2008–09, which have diminished in size but persist into 2018–19. The data show that automatic adjustments are slow at best. Preliminary estimates and forecasts for 2020-21 now expect large negative output gaps as a result of policies to contain the COVID-19 pandemic.

An aggregate demand shock

Figure 23.6 shows the adjustment to an output gap that is built into the ADπ/ASπ model. In the diagram, an unexpected fall in aggregate demand shifts the ADπ function to the left. A decline in exports or investment or government expenditure, or a change in the inflation target set for monetary policy, would have this effect on ADπ. Before this change, the economy was in equilibrium at full employment and potential output with inflation rate π^*, the central bank's inflation target.

Figure 23.6: Adjustment to a recessionary gap



A fall in ADπ to ADπ_1 creates a recessionary gap $(Y_1 - Y_P)$. The increase in unemployment and reduction in output cuts the rate of increase in wage rates and prices to π_1. Lower inflation leads the central bank to reduce its interest rate, supporting a move along ADπ_1 to Y_2.

In the short run, the fall in ADπ creates a recessionary gap $(Y_1 - Y_P)$. Since producers cannot cut costs per unit of output, they reduce employment and output to Y_1 to cut total costs. Unemployment rates rise. At Y_1 the goods market clears. It is a point on both the ASπ^* and ADπ_1 curves. Inflation has not fallen because of the short-run stickiness in wage rate increases and price increases.

As time passes there are opportunities to negotiate smaller wage rate increases and to reduce some planned price increases. The scale of these changes depends on the size of the output gap and the sensitivity of costs and prices to that gap. In the ASπ function the term $\gamma[(Y_1 - Y_P)/Y_P]_{t-1}$ defines this adjustment. ASπ shifts down to ASπ_1. As long as the central bank reacts to the output gap and lowers its overnight interest rate, expenditure can increase along ADπ_1 to Y_2 and offset some of the output gap.

In the real world, adjustments to changes in aggregate supply come from the effects of GDP gaps and unemployment on the wage agreements negotiated, formally and informally, in labour markets. A recessionary gap like that in Figure 23.6 lowers *wage rate increases*, lowers inflation, and allows monetary policy to support an increase in demand *along the ADπ_1 curve*. The output gap shrinks gradually but the inflation rate is below the central bank's target.

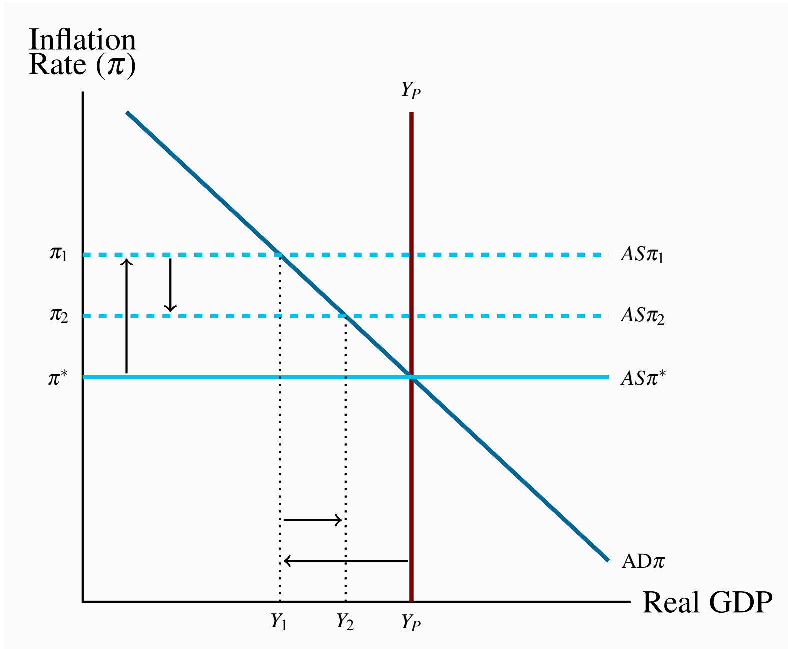
An inflationary gap results in the opposite process. Employment and output rise. Unemployment rates fall. Wage rate increases rise, inflation rises, and monetary policy reactions result in higher interest rates to reduce expenditure. In both cases, the eventual changes in wage rate agreements and rates of inflation together with the reaction of monetary policy to changes in inflation rates move the economy, over time, from short-run to long-run equilibrium. But the new equilibrium inflation rate would be higher than the Bank's target.

This adjustment process has important implications for the inflation targets set by monetary policy. It means prolonged changes in aggregate demand conditions require changes in the central bank's interest rate setting to defend the inflation target. It also means that, if the inflation target is cut, shifting the ADπ curve to the left, the economy will go through a recession, and perhaps a prolonged recession, while money wage rate agreements are renegotiated and price setting practices adjust to reflect the new inflation target. The time required for this adjustment is linked in a very important way to the independence and the credibility of the central bank. Canadian experience in the 1990s provides an interesting example. Recessionary gaps were persistent after the Bank of Canada adopted an inflation target of 2 percent at a time when current inflation rates were about 5 percent.

An aggregate supply shock

Figure 23.7 shows the model's internal adjustment to unexpected shift in ASπ. A sharp increase in crude oil and commodity prices, for example, enters the ASπ function as an increase in σ ($\Delta\sigma > 0$) in Equation 23.3. That pushes costs up and prices follow, increasing the inflation rate at least temporarily. Increases in indirect taxes like the HST or the taxes on alcohol and gasoline would have a similar effect. They increase the sellers' tax remittances to government, which sellers attempt to offset by increasing prices.

Figure 23.7: An aggregate supply shift



An increase in commodity prices ($\Delta\sigma > 0$) increase π as ASπ shifts up to ASπ_1. Inflation is higher than the Bank's π target and output falls, creating a recessionary gap. This is *'stagflation'*. The central bank faces a dilemma. It cannot defend its inflation target without increasing the gap. Alternatively it can wait until π falls back to π^*.

In Figure 23.7 the ASπ curve shifts up to ASπ_1. The inflation rate is higher and output falls creating a recessionary gap. These are the economic conditions often described as 'stagflation'. They pose a dilemma for the central bank.

Inflation is higher than its official target rate but trying to bring inflation down by raising the policy interest setting would only make things worse by raising the unemployment rate. Alternatively, cutting the policy interest rate to reduce the recessionary gap would shift ADπ to the right and validate the higher inflation rate. Doing nothing, on the assumption that the increase in commodity prices is a one-time jump, would mean living with inflation above the target and high unemployment until the rates of increase in wage rates and prices fall. Then ASπ would shift down to ASπ^*. The cost of waiting is the cumulative loss of employment and output during the adjustment process and the redistribution of income in favour of commodity producers if that is the source of the supply shift.

23.5 Monetary and fiscal policies

The inflation-output model sets clear roles for monetary and fiscal policies, at least in normal times. Monetary policy, which is embedded in the ADπ curve, is aimed at an inflation control target using a short-term interest rate instrument. Within the structure of the model this monetary policy simultaneously moderates business cycle fluctuations in output and supports equilibrium output at potential output. Fiscal policy can be expansionary, stimulating, or aimed at a government debt ratio target through budget balance control. The net tax rate in the government's budget provides automatic fiscal stabilization to moderate business cycle fluctuations.

Canadian monetary and fiscal policy followed this pattern from the early 1990s until the financial crisis of 2008. The Bank of Canada, in consultation with the Ministry of Finance has set inflation control targets for monetary policy since 1991. With the policy independence provided by a flexible exchange rate, the Bank uses the overnight interest rate as its policy instrument and intervenes in the market for overnight clearing balances. This keeps the overnight rate within the operating band it sets. An interest rate policy rule provides a useful description of the Bank's reactions to inflation rates that deviate from its control target.

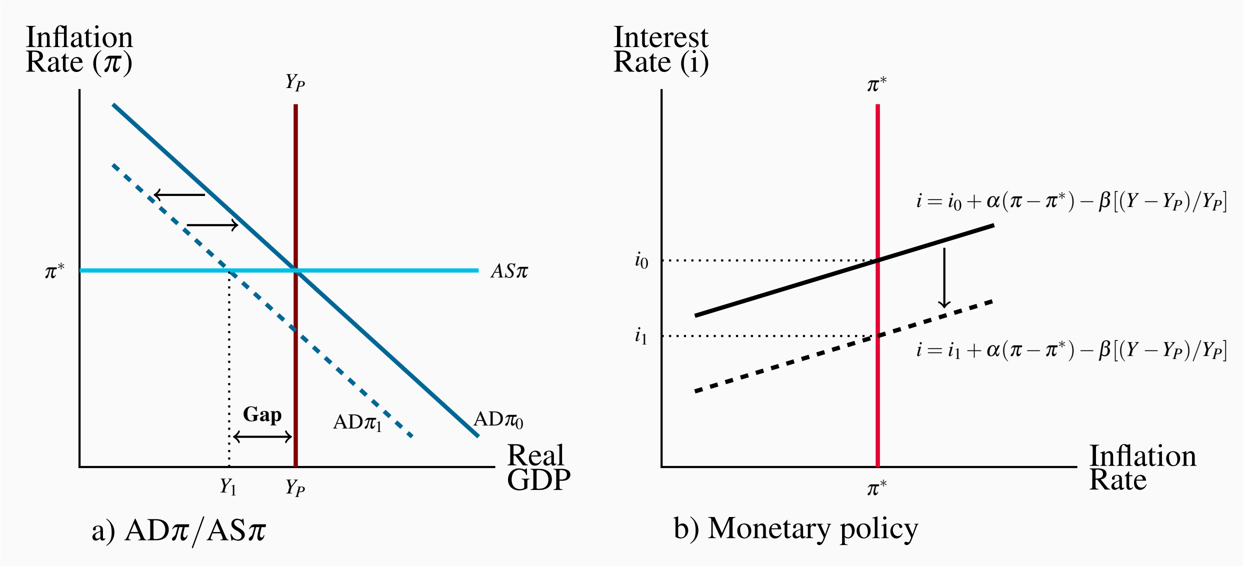
Monetary policy

In practical terms central banks set their short-term interest rate instruments based on their inflation rate control target and their assessment of current and future economic conditions. This approach is described by a policy rule for setting the overnight interest rate like that introduced in Chapter 22, namely:

$$i_{onr} = i_0 + \alpha(\pi - \pi^*) + \beta\left[(Y - Y_P)/Y_P\right] \tag{23.4}$$

Adding a term for the output gap means the central bank looks at both the inflation rate and the state of the economy in making its interest rate setting decisions. In terms of the ADπ/ASπ model, the output gap is a predictor of the way the inflation rate will move if current conditions persist. In times of recession, for example, the output gap term in the policy rule calls for a cut in interest rates even if the inflation rate is still π^*. This way the central bank pre-empts the fall in inflation that a recession would cause.

Figure 23.8: Central bank lowers overnight rate target



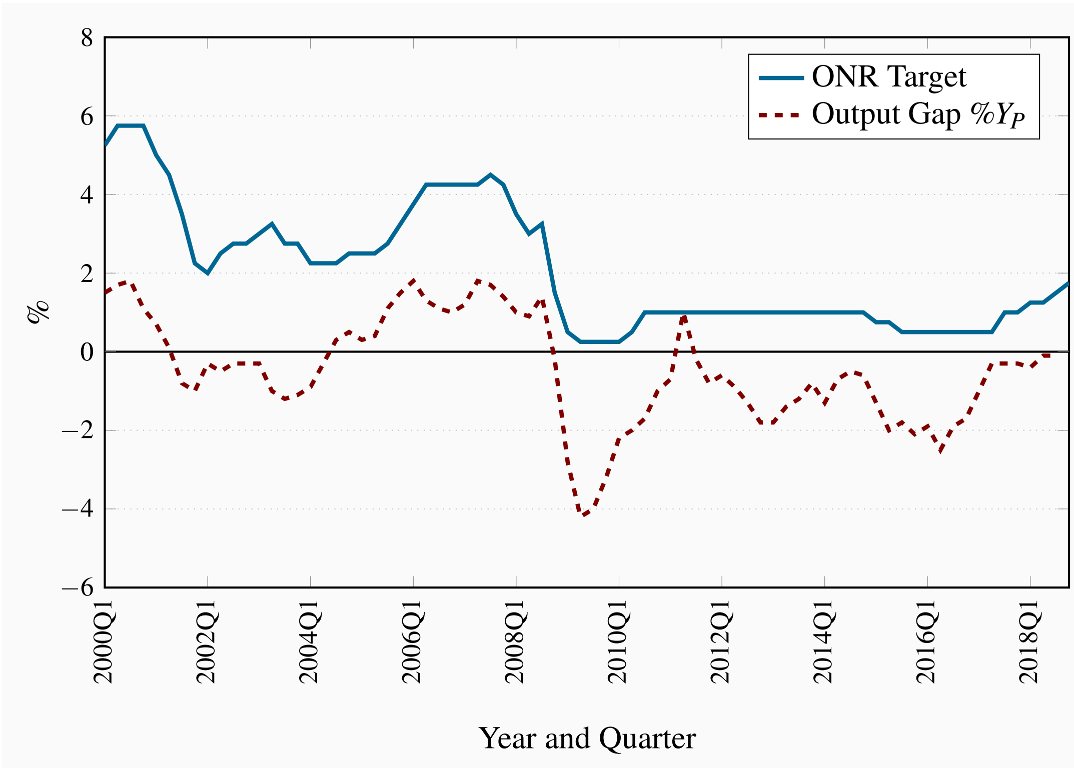
Initially, equilibrium is $Y_P\pi^*$ in a) with the Bank's overnight rate set at i_0, the neutral rate. Then $AD\pi$ falls to $AD\pi_1$ as a global recession reduces exports. GDP falls to Y_1 causing an output gap $Y_1 - Y_P$. The central bank reacts to the output gap with a cut in its overnight rate target to i_1 in b), to provide stimulus and increase $AD\pi$.

Figure 23.8 illustrates the central bank's policy response to a persistent fall in $AD\pi$. Lower $AD\pi$ creates a recessionary gap but does not immediately lower the inflation rate. The central bank sees the fall in $AD\pi$ as a lasting change in economic conditions that will reduce inflation below its inflation control target. It cuts its policy interest rate setting from i_0 to i_1 to increase $AD\pi$, shifting $AD\pi$ to the right, to reduce the output gap and downward pressure the gap puts on the inflation rate.

The Bank of Canada's cut in its overnight rate target by 75 basis points (3/4 of 1 percentage point) in December 2008 is a real world example of this monetary policy. In the press release announcing this rate cut the Bank explained that the outlook for the global recession was worse than earlier predicted. The prospects for Canadian GDP growth and inflation were weaker than previously thought. In these circumstances the cut in the overnight rate was intended to provide some offsetting monetary stimulus. In terms of Figure 23.8, the cut in the interest rate setting is the downward shift in the monetary policy rule. It was intended to pre-empt the shift from $AD\pi_0$ to $AD\pi_1$ at least in part. You can read the full text of the press release at: `http://www.bankofcanada.ca/2008/12/press-releases/bank-canada-lowers-overnight-rate-2/`

Figure 23.9 shows that the Bank of Canada's setting for its overnight interest rate has been related to the output gap over the 1995–2019 period. The interest rate setting rule described by Equation 23.4 calls for the Bank to lower its overnight rate as the output gap $(Y - Y_P)/Y_P$ is < 0 and increase its rate when the output gap $(Y - Y_P)/Y_P$ is > 0. Figure 23.8 gives one example of how this works in theory. It is indeed the pattern of overnight rate settings we see in the data plotted in Figure 23.9.

Figure 23.9: The ONR and the Output Gap



Source: Statistics Canada Table 10-10-0139-01, Bank of Canada, Indicators and author's calculations.

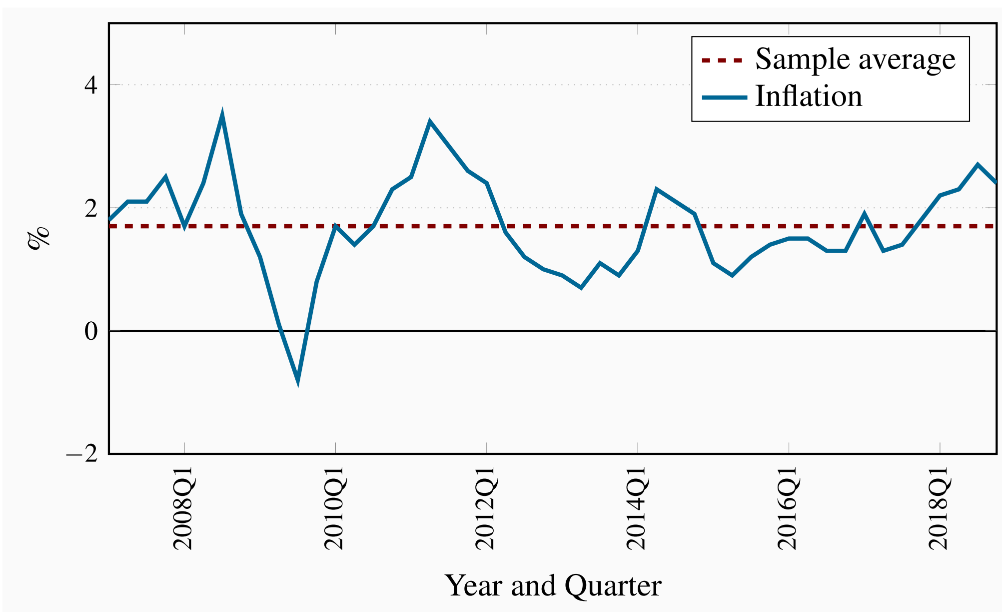
More examples of the Bank's reaction to an anticipated recession were the cuts in the overnight rate in 2015 following the drop in oil and commodity prices. This rate cut was intended to offset the negative effects of lower oil prices on the revenue, employment and investment expenditures of major oil and commodity producers. The economy was already working with a recessionary gap and that gap was growing in the first half of 2015. Then as economic growth increased after later 2017, reducing the output gap and pushing up the inflation rate, the Bank made successive increase in its overnight rate moving it up in steps from 0.5% to 1.75% in 2018Q4 and holding it there until 2019Q3.

Then in March and April, 2020, in response to government policies introduced to control the COVID-19 pandemic, the Bank moved to support the financial system and the economy. It lowered its overnight interest rate to 0.25 percent and increased open market operations. It also launched new asset purchase programs to support funding for business and the Government of Canada debt market. For more details on these policy moves see the Bank of Canada press releases of March 27, 2020 and April 15, 2020. https://www.bankofcanada.ca/2020/04/fad-press-release-2020-04-15/ and https://www.bankofcanada.ca/2020/03/press-release-2020-03-27

Changing overnight interest rates in response to changes in the output gap is only one aspect of the interest policy rate rule. The underlying argument is that output gaps are indicators of future changes in inflation rates. This is consistent with the ADπ/ASπ model, which explains changes in the inflation rate that follow from persistent output gaps. The ASπ function also provides for changes in the inflation rate as a result of other economic changes such as changes in the prices of raw materials or changes in indirect taxes. The Bank might change its overnight rate setting in response to these changes in inflation that are not driven by the output gaps.

Figure 23.10 shows the success of the Bank of Canada's monetary policy over the 2007Q1–2018Q3 period. The inflation rate as measured by the annualized rate of change in the Consumer Price Index has averaged 1.8 percent. This is close to the midpoint of the Bank's inflation control target range. Monetary policy cannot eliminate all variations in the inflation rate. Many changes come and go before the underlying causes are recognized. Furthermore, it takes considerable time for changes in the interest rate to work their way into the economy and affect inflation. Recognizing the transitory nature of short-term fluctuations in the inflation rate and the time lags in the effects of interest rate changes the Bank aims to hit its inflation rate target within a time frame of six to eight quarters of a year.

Figure 23.10: Canadian Inflation: Quarterly at Annual Rates, 2007Q1–2018Q3



Source: Bank of Canada: Total CPI inflation and author's calculations.

The data plotted in Figure 23.10 show the volatility of the annualized quarterly inflation rates but the average rate over the period is indeed 1.8 percent, just below the Bank's 2.0 percent target. Over the same period the economy went through both recessionary and inflationary gaps, with a small recessionary gap, -0.5 percent of Y_P, on average. If the period following the financial crisis of 2008 is excluded the average output gap from 1995 to 2008 is -0.2 percent with a range of -2.1

percent to +2.4 percent.

While success with inflation control did not eliminate the business cycle in real GDP and employment, economic performance was consistent with the argument of the ADπ/ASπ model. From 1995 to 2019 the inflation rate fluctuated around the 2 percent target and output was, on average, close to estimates of potential output.

The financial crisis of 2008 changed the economic fundamentals, the focus of monetary and fiscal policies, and the economic performance of the economy. This more recent experience is covered after a discussion of recent fiscal policy in terms of the ADπ/ASπ model.

Fiscal policy

Under normal conditions in the ADπ/ASπ model with a flexible foreign exchange rate, monetary policy dominates discretionary fiscal policy. Monetary policy sets the interest rate that will give the ADπ needed to get the inflation rate target at potential output. A change in discretionary fiscal policy would change the conditions on which the Bank's interest rate is set. To defend its inflation rate target the Bank would react, changing its interest rate setting to offset any new fiscal stimulus or restraint. Monetary policy aimed at inflation control deliberately crowds out the effect of fiscal policy on ADπ through the interest rate-exchange rate links in the monetary transmission mechanism.

For example, in 2006 and 2007 the Government of Canada cut the GST by a total of 2.0 percentage points. The Bank of Canada raised its overnight interest rate from 3.25 percent to 4.50 over the same time period. At the time the economy was also working at an inflationary gap. The increase in interest rates offset the fiscal stimulus from the GST cut by raising costs of current and future credit.

However, fiscal policy does have important roles to play even when the central bank successfully targets inflation and potential output:

1. Provide automatic fiscal stabilization; and
2. Control, and perhaps reduce, the ratio of the public debt to GDP.

Fiscal policy is implemented through the government's budget. The budget balance function provides the framework. Elaborating on the basic government budget of Chapter 19 to recognize the importance of the public debt gives:

$$BB = tY - G - iPD \tag{23.5}$$

The term iPD is the interest paid on the government bonds, PD, issued in the past to finance budget deficits. G represents *program spending*. $G + iPD$ is total government spending.

The fiscal policy instruments are:

1. The net tax rate, t, set in the budget plan, which will also be a component of,
2. The **primary budget balance**, $PBB = tY - G$, which *excludes* interest payments on the public debt.

The Department of Finance, *Fiscal Reference Tables, 2019* (`www.fin.gc.ca`) report the Government of Canada paid interest (iPD) of $23.3 billion in the fiscal year 2018–19. That was an amount equal to 2.3 percent of its outstanding interest bearing debt and 6.7 percent of total government expense ($G + iPD$).

Government tax and expenditure programs linked to GDP provide *automatic* fiscal stabilization, as explained in Chapter 19. In addition, with a monetary policy consistent with equilibrium at potential output, *discretionary* fiscal policy can aim to control the **public debt ratio (PD/Y)** by managing the government's **primary budget balance**. This is the fiscal policy objective that has caused serious debates in both the US and Europe, and in the 2015 and 2019 federal election campaigns in Canada.

> **Public debt ratio (PD/Y)**: the ratio of government debt to GDP.
>
> **Primary budget balance ($PBB = tY - G$)**: the difference between net tax revenue and government program expenditure. It excludes interest payments on the public debt.

Current interest payments on the public debt are the result of past issues of government bonds and the average of coupon rates on those bonds. This expenditure cannot be changed in the current budget but can be controlled going forward as current budget balances reduce or increase the public debt.

Automatic stabilization

Automatic fiscal stabilization comes from the net tax rate t in the budget function. The net rate changes government revenues and transfers and the budget balance automatically when GDP changes. These changes in the government budget balance reduce the size of the expenditure multiplier and the size of the change in GDP caused by shifts in $AD\pi/AS\pi$ conditions. The budget function:

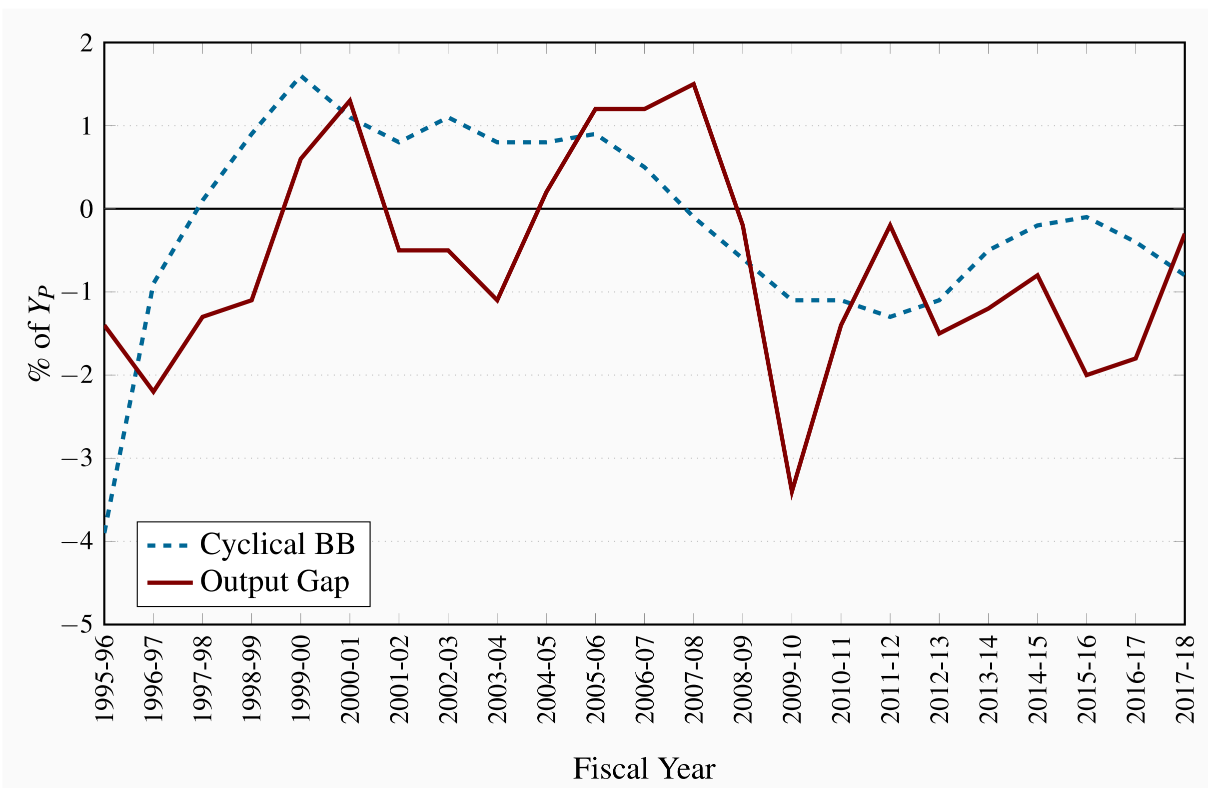
$$BB = SPBB + t(Y - Y_P) - iPD \tag{23.6}$$

separates the *cyclical component* $t(Y - Y_P)$ of the budget balance BB from the structural primary budget balance $SPBB$. The structural primary budget balance evaluates the government's fiscal program *net of interest payments on the public debt* at Y_P. The cyclical component is the effect of an output gap on the actual budget balance.

Setting the net tax rate that will provide automatic stabilization is a simple mechanical question but a difficult economic and political question. A higher net tax rate provides more stabilization. It involves the design of the tax system and transfer payment programs like employment insurance, social assistance programs and subsidy programs where revenues collected and expenditures made vary inversely to economic conditions. The types of taxes to be used and the bases to which those taxes are to be applied generate heated debates. The design and eligibility requirements for transfer benefits are also controversial. There are also strong differences of opinion over the economic impacts of different types and rates of taxes and transfers. These issues have been important to recent election campaigns in Canada and the US and dominate debates about budget deficits and public debt control.

We will leave the details of those debates to specialists in public finance and ask a question that is somewhat easier to answer. Have changes in the cyclical balance of the Canadian federal government budget provided automatic stabilization? To answer this question, Figure 23.11 plots the output gap and the cyclical budget balance over the 1995–2018 time period. These data show a pattern of cyclical budget deficits and surpluses that coincide with recessionary and inflationary gaps to provide automatic fiscal stabilization.

Figure 23.11: Output Gaps and the Government of Canada Cyclical Budget Balance, 1995–96 to 2017–18



Source: Department of Finance, Fiscal Reference Tables, 2018: Table 17, Bank of Canada Indicators 2018 and author's calculations.

For most years in the period covered by Figure 23.11 the absolute magnitude of the cyclical budget balance is noticeably less than that of the output gap. That pattern changes after 2008 when expansionary fiscal policy, the *'Economic Action Plan'* of 2009, was introduced to fight the recession that followed the 2008 financial crisis. Although this was clearly a discretionary policy change, the Department of Finance reported it in its *Fiscal Reference Tables, 2011*, as a component of the cyclical balance, excluding it from its structural budget balance estimates, and continues this practice with additional revisions in the *Fiscal Reference Tables, 2019*. This is an interesting treatment of this discretionary stimulus program that was extended into 2011 and 2014 and continued to feature prominently in government advertisements. As a result, the 'automatic stabilization' illustrated in Figure 23.11 is overstated in the 2009–2014 period and the reported structural budget balance is increased. Furthermore, in the 2013–2015 period the changes in cyclical budget balance reflect a shift to fiscal austerity driven by the Harper Conservative government's commitment to achieve a **balanced budget**, $BB = 0$, by 2015 even in a time of economic recession.

The Liberal government elected in 2015 offered an alternative approach. Government budget deficits for at least a few years based on infrastructure expenditure. The objective was to stimulate economic growth, reduce recessionary income gaps and unemployment by increasing aggregate demand and productivity. As economic conditions deteriorated as a result of falling oil and commodity prices budget deficits were larger than initially proposed, undermining the original time frame set for achieving a balanced budget. Control of the **public debt ratio** became the new constraint on fiscal stimulus and the budget deficit.

Setting a target for the public debt ratio defines both the size of the primary budget balance and the budget deficit the government will run. It also defines the scale of the fiscal expansion or austerity the government has to work with.

Controlling the debt ratio

The *public debt ratio* is usually defined as the outstanding public debt as a percentage of annual nominal GDP, namely PD/Y. Explaining the budget balances required to control and reduce the *public debt ratio* as time goes by is a bit more complicated.

Start with the link between the outstanding debt and the government budget balance. At any point in time the outstanding public debt is the sum of all previous budget deficits and surpluses. The annual change in the outstanding public debt, PD, is equal to the annual budget deficit or surplus. A budget surplus, $BB > 0$, will reduce the outstanding public debt as some of the outstanding debt (bonds) will mature and new debt will not be issued. On the other hand, a budget deficit, $BB < 0$, will be financed by issuing new bonds that add to the outstanding debt. In summary, the change in the public debt from one year to the next, ΔPD, is equal to the inverse of the budget balance, namely, $-BB$.

The change in the *public debt ratio* from one year to the next is determined by two conditions namely:

1. The change in the debt, ΔPD, which is the change in the numerator of the ratio and

2. The change in GDP, ΔY, which is the change in the denominator of the ratio.

Recall that the government's budget balance function is:

$$BB = tY - G - iPD.$$

The net tax rate t, and government program expenditure G are set in annual budgets. The third component, iPD, is determined outstanding public debt and the *average interest rate on that debt*. As a result it is only the primary budget balance ($PBB = tY - G$) that can be changed in annual budgets.

Furthermore, if the primary budget is balanced, ($PBB = tY - G = 0$) interest payments on the outstanding debt will still increase the debt by iPD. For example, if annually the rate of interest paid on the debt is 4%, ($i = 0.04$), the debt will grow by 4% a year. This increase in the public debt will increase the debt ratio unless there is an offsetting increase in GDP.

But nominal GDP also grows from year to year, except in times of recession. As a result if the average annual growth in nominal GDP, ($\Delta Y/Y_{-1}$), is equal to the interest rate on the public debt, we say 4%, ($r = 0.04$), the public debt ratio would be constant from year to year. As a result, the budget balance relative to GDP required to keep the debt ratio stable depends on the *difference between the annual rate of interest paid on the debt and the growth rate of nominal GDP*.

The Department of Finance "*Fiscal Reference Tables 2018*" provide data for an empirical example. In Canada from 2015–2018, the annual interest rate on the outstanding public debt averaged 2.23%. The average annual growth rate of nominal GDP in the same period was 3.43%. The GDP growth rate was greater than the interest rate by 1.2%, which would reduce the debt ratio by 1.2/

Table 23.1 provides a numerical example of a change in the Government of Canada's debt ratio based on a small sample of recent data from the Department of Finances Fiscal Reference Tables 2016. Columns (2), (3) and (4) show the interest rate, growth rate and public debt ratio used to estimate the primary budget balance required to stabilize the public debt ratio at its fiscal year end value of 31.3 percent.

Based on these conditions, with the interest rate on the public debt, $i = 3.2$ percent, less than the growth rate in nominal GDP, $g = 4.5$ percent, the debt ratio would have been constant even with a small primary budget deficit, $PBB < 0$, of about 0.04 percent. As it turned out, the government's actual primary budget surplus reduced the debt ratio from 31.3 percent in 2013–14 to 29.6 percent in 2014–15. Strong growth in nominal GDP was important for this outcome.

Table 23.1: A numerical example of the primary budget balance required for a constant debt ratio

(1) Time Period	(2) Interest rate on public debt i	(3) Growth rate of nominal GDP g	(4) Public debt ratio PD/Y	(5) PBB/Y required for $\Delta(PD/Y)=0$	(6) Observed PBB/Y	(7) Observed $\Delta(PD/Y)$
2013–2014	0.027	0.046	0.331	-0.006	0.008	-0.012
2017–2018	0.022	0.039	0.306	-0.005	0.002	-0.003

Source: Department of Finance, *Fiscal Reference Tables, 2018*, Tables 13, 15, and 17, Statistics Canada Table 36-10-0122-01 and author's calculations.

The current Government of Canada is not the first to identify the debt ratio as a constraint on fiscal policy. In 1994 the Government of Canada announced *A New Framework for Economic Policy* that focused fiscal policy on deficit control and debt ratio reduction. See esp. p. 82: https://www.scribd.com/document/340386383/NFrmrkEcPol-e

When this policy announcement was made the public debt stood at 68 percent of GDP. The primary budget balance was in deficit and interest payments on the public debt absorbed about 35 percent of government revenue. In the years that followed, the public debt ratio declined steadily to a low 29 percent in 2008–09. Interest payments on the debt fell to less than 13 percent of revenue. Fiscal policy was on track to meet the debt ratio target of 25 percent set in the 2004 Budget. Significant primary surpluses were the major fiscal contributor to this decline in the ratio.

Canadian success with debt ratio control and reduction from 1995 until 2009 has been cited as an example for countries currently facing very high debt ratios. But it may not fit with current economic conditions in those countries. Canada's fiscal adjustment was made easier by a substantial easing of monetary conditions in terms of both interest rates and exchange rates and strong growth in the US economy, which all supported a large increase in exports and growth in Canadian GDP. This economic environment is not available to the countries that now face a debt crisis.

Even with favourable conditions, the Canadian adjustment was costly. There was a cumulative loss of output from an output gap that lasted from 1995 to 1999 that was at times as large as 2 percent of Y_P. Unemployment rates reached 9 percent and low rates of capacity utilization in industry reduced profitability and investment in new capacity and technology.

These and the negative fiscal effects of the earlier monetary restraint following the shift to inflation targeting are now sunk costs. In the years that followed monetary and fiscal policies worked to deliver stable inflation and declining public debt ratios until the crisis and recession of 2008–09.

23.6 Recession, disinflation and deflation

The financial crisis of 2008 and the recession that followed uncovered the limitations of monetary policy and renewed the case for fiscal stimulation. The assignment of monetary policy to an inflation target at Y_P and fiscal policy to debt ratio control did not work in the new economic conditions. Moreover, the depth of the recession and the collapse of the U.S. banking and financial sectors led to **disinflation** and raised fears of **deflation** as a result of rising real interest rates and output gaps. The Great Depression of the 1930s was an historical example. The Japanese experience starting in the 1990s with zero interest rates and a continuing slump was another.

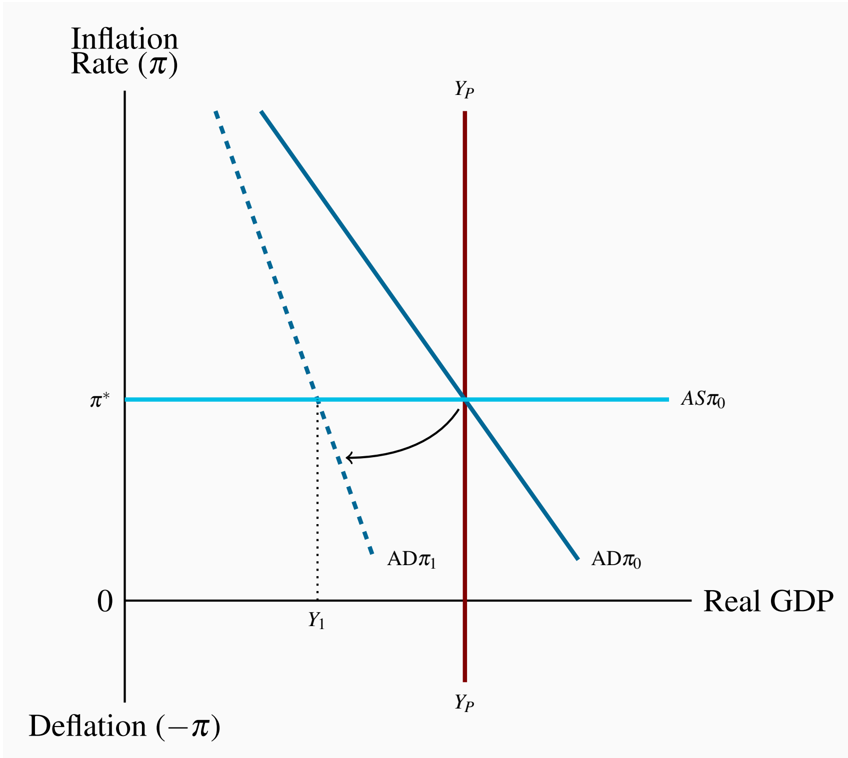
Although these experiences are historic and economic conditions and policy have normalized, Stephen Poloz, the Governor of the Bank of Canada frequently speaks about the challenges, limitations and risks faced by the Bank of Canada in formulating monetary policy. The full text of that speech on these topics is available on the Bank of Canada website.

Although the risk of deflation has receded the issues raised by *disinflation* remain whenever central banks have reached the zero lower bound and cannot provide monetary stimulus through interest rate cuts. A rethinking of macroeconomic policy was necessary.

> **Disinflation**: a persistent fall in the inflation rate.
>
> **Deflation**: a persistent fall in the general price level.

Some of the issues involved can be explained with the basic $AD\pi/AS\pi$ model. In Figure 23.12 a financial crisis shifts $AD\pi$ by undermining the interest rate-expenditure link in the monetary transmission mechanism. Financial risk and uncertainty increase and demands for liquidity increase. Banks are reluctant to lend and businesses are reluctant to borrow. The $AD\pi$ curve shifts left and pivots to show the reduction in financing and the fall in the interest sensitivity of expenditure. Equilibrium shifts to π^*Y_1 with a recessionary gap.

Figure 23.12: Recession, disinflation & monetary policy



A financial crisis undermines monetary transmission mechanism. ADπ falls but central bank interest rate cuts cannot offset the fall in π. At zero lower bound deflation raises real interest rate and the output gap continues to grow.

The recessionary gap triggers two reactions. The inflation rate falls and the central bank cuts its overnight interest rate to offset the fall in inflation and output. But the steeper ADπ_1 function reflects the disruption to financial markets. Nominal interest rate cuts have limited the effects on expenditure. To prevent a rise in the real interest rate $(i - \pi)$ the Bank must cut its nominal policy rate by more than the fall in inflation.

But the central bank cannot lower its rate below zero – it hits the lower bound. *If the output gap persists and pushes disinflation into deflation, rising real interest rates cut expenditure and increase the gap. The* ***ADπ function would be kinked backward at*** *$i_{onr} = 0$ to capture these negative effects of disinflation and deflation*. Falling inflation rates reduce aggregate expenditure and aggregate demand along the ADπ curve.

Fortunately economic conditions did not deteriorate badly enough during the 2008–09 financial crisis to trigger deflation:

> *"It appears today that the world will likely avoid a major deflation and thus avoid the deadly interaction of larger and larger deflation, higher and higher real interest rates*

> *and larger and larger output gaps."*
>
> Blanchard, O. et al. (2010). Rethinking Macroeconomic Policy. ***IMF Staff Position Note***, SPN/10/03

Nevertheless, faced with the financial crisis and recession, major industrial countries shifted their policy programs. Cutting interest rates to defend inflation targets was the first policy initiative. The Bank of Canada cut its target overnight interest rate from 4.25 percent in late 2007 to a record low 0.25 percent in early 2009 and held the rate at that level until mid-2010. In the US the Federal Reserve cut its federal funds rate, the policy rate that corresponds to the Bank of Canada's overnight rate, from 5.25 percent in late 2007 to 0.0 percent in late 2008. It was still at that rate in late 2014. These interest rate cuts reduced policy rates to the zero lower bound and exhausted the potential for monetary stimulus with that policy instrument.

Monetary policy in the US turned then to instruments used before it adopted interest rate setting to get an inflation target. The first was increased "moral suasion," an increase in communications with financial market participants, including *conditional statements about the future path of policy rates*, to emphasize the central bank's longer-term support for markets and its actions to promote stability. More directly the banks were urged to maintain their lending operations. The second was "**quantitative easing**," and in the case of the US, an even more extensive "**credit easing**." "Quantitative easing" extends the use of open market operations described above. The central bank purchases a broader range of financial assets to expand its balance sheet to increase substantially the monetary base and cash positions of the banks. In other words, the objective is to increase the quantity of cash reserves in the banking system directly.

> **Quantitative easing**: a central bank purchase of financial assets to increase its asset holdings and the monetary base.
>
> **Credit easing**: the management of the central bank's assets designed to support lending in specific financial markets.

Open market operations usually involve central bank purchases of short-term government bonds. The US Federal Reserve went beyond this, introducing three sets of policy tools:

1. lending to financial institutions,
2. providing liquidity directly to other key credit markets by buying highly rated commercial paper, and
3. buying longer-term securities, including mortgage backed securities.

Using these tools increases the size of the central bank's balance sheet and changes the structure of central bank asset holdings. Quantitative easing is measured by the impact on the quantity of bank reserves. Credit easing is measured by the wider variety of loans and securities the central bank willingly holds on its balance sheet. A purchase of these assets puts cash directly into specific markets rather than feeding it through banks and bank lending. Both are intended to increase lending to businesses and households in times when very low, near zero, interest rates alone are not working.

The effect of three rounds of quantitative easing by the US Federal Reserve was to raise the monetary base in the US from about $800 billion in to $2.62 trillion in October 2012. While this monetary stimulus did not offset the full effects of the 2008-09 recession, it was credited with avoiding a US deflation. A version of this policy action was used in Japan earlier in the decade after the Bank of Japan had lowered its borrowing rate to zero and wanted to provide further economic stimulus. In terms of the ADπ/ASπ model the intent was to support demand and shift the ADπ curve to the right or least avoid deflation if recessionary gaps persisted.

The Bank of Canada also developed plans for quantitative easing that were not used as economic conditions improved more quickly here than in some other countries. You can see the Bank's explanation of these policies in its Monetary Policy Report, April 2009 pp. 24-28, at `http://www.bankofcanada.ca/wp-content/uploads/2010/03/mpr230409.pdf`.

These monetary policies were supported by discretionary fiscal policies. With interest rates a zero, crowding out was not a concern and increased ADπ was the objective. The earlier focus on deficit and debt ratio control was set aside to provide aggregate demand stimulus. In Canada the federal government's '***Action Plan***' included direct support for infrastructure projects and tax incentives for certain household expenditures. In terms of budget data the federal government structural budget shifted from a surplus of $3.8 billion in fiscal 2007–08 to a deficit of $10.6 billion in 2009–10. But as noted earlier, the 'Action Plan' stimulus was treated as a 'cyclical' budget component. Thus the change in the actual federal budget balance from a surplus of $9.6 billion in fiscal 2007–08 to a deficit of $55.6 billion in fiscal 2009–10 is probably a better measure of the fiscal stimulus that addressed the recession. However, this stimulus was relatively short lived as the government returned to its earlier deficit and debt control focus and left monetary policy to provide sustained support for economic recovery. The monetary/fiscal policy mix remained that way in mid-2015 leading up to the federal election in the fall of that year.

The 2015 federal government led by Prime Minister Justin Trudeau shifted from a balanced budget fiscal plan to budgets designed to promote economic growth based on major infrastructure investment and modest tax cuts financed by modest budget deficits. This fiscal stimulus together with continued monetary stimulus from the Bank of Canada shifted policy mix from fiscal austerity/monetary stimulus to fiscal and monetary stimulus combined. Management of the public debt ratio emerged as the constraint on fiscal policy, replacing the previous government's focus on balanced budgets.

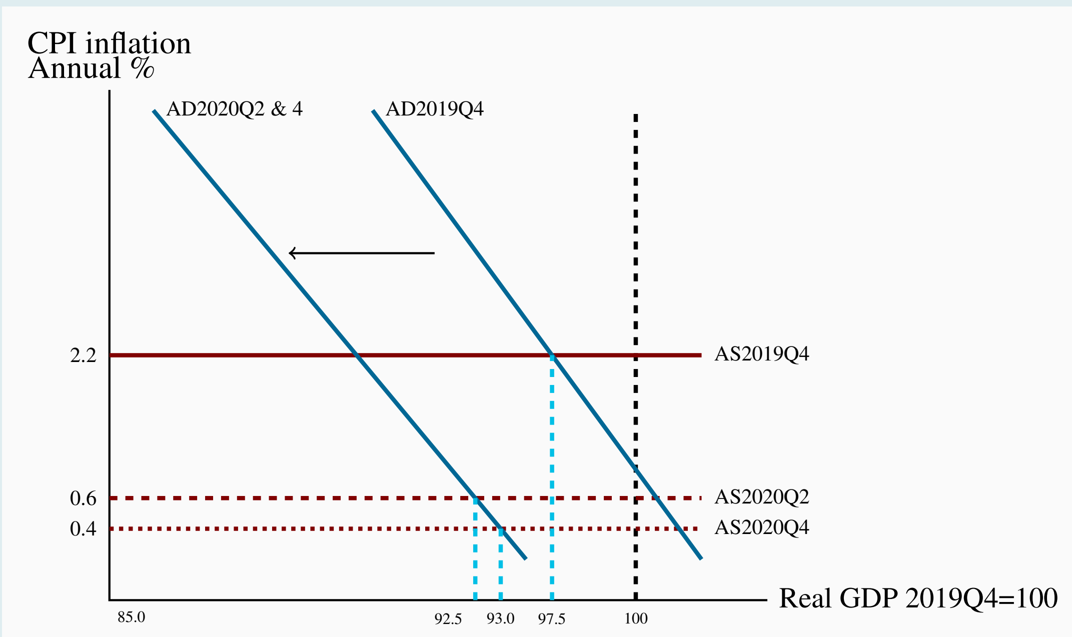
The federal government re-elected in 2019 planned to continue this approach to fiscal and budgetary policy but these plans were disrupted by the COVID-19 pandemic. The budget 2020-21 presentation was delayed when the House of Commons shut down in late March as part of the plan to control the spread of the pandemic.

The Application and Example boxes below illustrate the effects of policies to contain the spread of the COVID-19 pandemic, in an AD/AS diagram model, and an example of the basic algebra AD function.

Application Box 23.1: The Impact of COVID-19 policies in AD/AS

AD/AS model can be used to illustrate the changes in economic conditions caused by the COVID-19 pandemic and policies to contain its spread. Estimates of changes in inflation and real GDP caused by changes in AD and AS are the result of business shutdowns and physical distancing. However, it is important to emphasize that the diagram illustrates the estimates published by the Parliamentary Budget Office. Deriving these estimates requires a fully specified and calibrated model that we don't have.

Figure 23.13: Changes in AD/AS and effects on Real GDP and Inflation



The PBO Report, Figure 2.1 shows estimates of real GDP at annual rates in the 2020Q1 will be 2.5 percent lower than in 2019Q4 and 7.0 percent lower in 2020Q4, after recovering slightly from a low in 2020Q2. The corresponding 2020 inflation rates, measured by the CPI are 0.6 and 0.4 percent compared to 2.1 percent in 2019Q4.

In the diagram, 2019Q4 inflation and real GDP, by AD_{2019Q4} and AS_{2019Q4} are 2.1% and 97.5. The initial impacts of COVID-19 policy shift AD to $AD_{2020Q2\&4}$ and AS to AS_{2020Q2}. Real GDP falls by 7.5% relative to 2019Q4 and inflation drops to 0.6%. Then by 2020Q4 the AD curve is unchanged but the drop in inflation to 0.4% comes with a small increase in real GDP of 0.5%.

Example Box 23.1: The Algebra of the ADπ curve

The ADπ curve is based on:

1. An AE function that includes the effect of interest rates on expenditures, and
2. Central bank monetary policy that sets interest rates to defend an inflation rate target.

The AE function is:

$$\text{AE} = A_0 - vi + [c(1-t) - m]Y \tag{1}$$

In this equation A_0 is autonomous expenditure, v measures the impact of a change in the interest rate, i, on expenditure, and $[c(1-t) - m]$, the marginal propensity to spend. Using the equilibrium expenditure condition $Y = \text{AE}$ gives:

$$Y = A_0 - vi + [c(1-t) - m]Y$$

$$Y = (A_0 - vi)/(1 - c(1-t) + m) \tag{2}$$

The central bank sets interest rates to achieve a *target inflation rate* and reacts to changes in the inflation rate by changing short-term interest rates according to:

$$i = i_0 + \beta(\pi - \pi^*) \tag{3}$$

In this equation i is the Bank's interest rate instrument, π^* is the Bank's inflation rate target and π is the actual inflation rate. With $\beta > 1$ the central bank's response to a change in the inflation rate changes the nominal interest rate enough to change the real interest rate. The monetary policy objective is to keep inflation at the target value π^*.

Substituting the interest rate determined by monetary policy in (3) into the equilibrium expenditure condition (2) gives an aggregate demand curve that includes the inflation rate:

$$\text{AD}\pi = (A_0 - v[i_0 + \beta(\pi - pi^*)])/(1 - c(1-t) + m) \tag{4}$$

This aggregate demand curve is labeled ADπ to distinguish it from the traditional P, Y aggregate demand curve. ADπ is the relationship between *real output and inflation*.

A numerical example illustrates this ADπ curve. Suppose

$$\text{AE} = 200 + 0.5Y - 10i,$$

and the central bank has an inflation target $\pi^* = 1.0$ percent. The bank estimates that a real interest rate of 2.0 percent is needed for equilibrium at potential output. It sets the nominal interest rate $i_0 = (r + \pi^*) = 3.0$ and reacts to *transitory* departures from its inflation target according to the policy rule:

$$i = 3.0 + 1.5(\pi - 1)$$

Then by substitution,

$$\begin{aligned} \text{AE} &= 200 + 0.5Y - 10[3.0 + 1.5(\pi - 1.0)] \\ \text{AE} &= 185 - 15\pi + 0.5Y \end{aligned}$$

In equilibrium $Y = \text{AE}$ giving:

$$Y = (185 - 15\pi)/((1 - 0.5)) = 370 - 30\pi \tag{5}$$

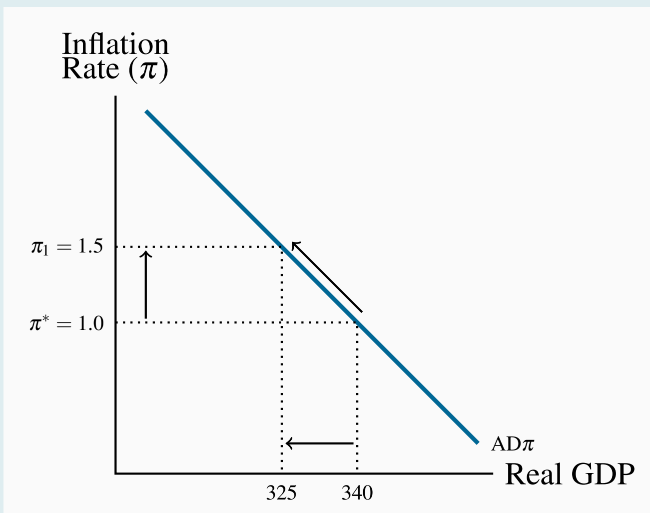
The aggregate demand curve:

$$\text{AD}\pi : Y = 370 - 30\pi$$

defines a negative relationship between the inflation rate and equilibrium real GDP.

When the inflation rate is equal to the monetary policy target of 1 percent ($\pi = \pi^*$), the equilibrium level of real GDP and demand is 340. If the inflation rate were to rise to 1.5 percent, which exceeds the target by 0.5 percent, the Bank's reaction would be to raise the interest rate by $[1.5(1.5 - 1)] = 0.75$ percent. This would lower aggregate expenditure by $(10 \times 0.75 = 7.5)$ and equilibrium real GDP by $(7.5 \times \text{multiplier} = 15)$ to 325, moving up along the ADπ curve. A fall in inflation from the target would bring the opposite reaction. This scenario is depicted in Figure 23.14.

Figure 23.14: The ADπ curve



Next

This chapter provides the basic model for the analysis of macroeconomic performance and policy that integrates the design, instruments and objectives of modern monetary and fiscal policies. Inflation, deflation, public debt ratios and business cycles are explained. But two issues remain. One is international macroeconomics and the importance of exchange rate policy for the effectiveness of monetary and fiscal policies. The second is economic growth, growth theory and growth accounting. These are covered in the next two chapters.

Key Concepts

The ADπ curve describes the relationship between equilibrium real GDP and the inflation rate.

Monetary policy that sets an **inflation rate target** and uses the **short-run interest rate** as a policy instrument is integral to the ADπ curve.

The **production function** $Y = A \times F(N,K)$ determines the output of goods and services.

In the long run, labour market equilibrium determines the equilibrium level of employment and **potential output** Y_P.

In the short run, employment and output fluctuate in response to changing economic conditions. However, there is **no trade-off between inflation rates and output**.

The **short run ASπ** function describes the lagged adjustment to the inflation rate in reaction to output gaps.

The ASπ function is based on the **price setting decisions** of producers in imperfectly competitive markets.

The **equilibrium inflation rate** is the inflation control target set by the central bank.

Business cycles in output and employment caused by persistent ADπ and/or ASπ shifts create output gaps.

Persistent output gaps lead to an **internal adjustment** based on **changes in the rate of increase in money wage rates and changes in the rate of inflation**.

A **monetary policy rule** describes the central bank's setting of its interest rate policy instrument as reactions to inflation rates that differ from its inflation control target and to output gaps.

Fiscal policy controls government budget balances to control or reduce the size of the **public debt ratio**.

The **primary budget balance** is the **fiscal policy instrument**.

The arithmetic of changes in the public debt ratio links the **change in the debt ratio** to the primary budget balance and the difference between the interest rate on the debt and the growth rate of GDP.

Over the 1995–2008 period Canadian monetary and fiscal policies effectively **controlled the Canadian inflation rate and reduced the federal government debt ratio**.

The financial crisis of 2008 and the recession that followed in 2008–09 demonstrated the limits of monetary policies based on interest rates and fiscal policies focused on debt ratio control. In Canada, a combination of monetary and fiscal stimulus reduced the fall in output and employment. In other countries, a broader set of monetary policy instruments and fiscal austerity have yet to solve employment and public debt issues.

Exercises for Chapter 23

Exercise 23.1 Use an AD_π/AS_π diagram to illustrate an economy at equilibrium with real GDP at Y_0 and an inflation rate of π_0. In this diagram, show how a permanent increase in exports would affect the equilibrium inflation rate and the equilibrium level of GDP if the central bank did not react and change its monetary policy.

Exercise 23.2 Suppose the central bank reacted to defend its inflation target from the effect of the increase in exports in Exercise 23.1. Use an ADπ/ASπ diagram to show the change in the equilibrium interest rate setting and real GDP you would observe.

Exercise 23.3 Suppose opportunities for investing in high tech applications boost aggregate demand in the short run, and aggregate supply in the long run. Using ASπ and ADπ curves with equilibrium at potential output, show why output might rise in the long run without much of an increase in inflation.

Exercise 23.4 Suppose a new round of labour negotiations results in a higher average rate of increase in money wage rates for the next three years. Illustrate and explain how this would affect short-run aggregate supply conditions and the ASπ curve.

Exercise 23.5 Draw an aggregate supply and demand curve diagram to show an economy in short-run equilibrium at potential output. Suppose a wide-spread recession reduces incomes in foreign countries, leading to reduced demand for exports. Illustrate and explain how this would affect the short-run equilibrium Y and π in your diagram.

Exercise 23.6 Suppose central banks have reduced their policy interest rates to the lower bound to fight a deep and prolonged recession. Use a diagram to show how either a reduction in the inflation rate, or deflation, would change the slope of the ADπ curve. Would cuts in nominal money wage rates and further reductions in the inflation rate reduce the recessionary gap when the central bank is constrained by the lower bound on its interest rate?

Exercise 23.7 In the two years before 2008 the Canadian federal government reduced the GST from 7 percent to 5 percent. Use an ADπ/ASπ/Y_P diagram to illustrate and explain the effects of this tax change on equilibrium output and inflation. If the economy was in equilibrium at Y_P and the target inflation rate π^* before the tax cut, what monetary policy action, if any, would the central bank make to maintain those equilibrium conditions after the tax cut? What short-run net benefit, if any, would households and businesses realize as a result of the cut in the GST?

Exercise 23.8 Define the 'public debt' and explain why and how it might increase from one year to the next.

Exercise 23.9 Would it be possible for the ratio of the public debt to GDP (PD/Y) to fall even if the government's primary budget is in deficit? Explain your answer.

Exercise 23.10 Would it be possible for the ratio of public debt to GDP (PD/Y) to rise even if the government's primary budget balance is in surplus? Explain your answer.

Exercise 23.11 *Optional*: Suppose the central bank's monetary policy sets the interest rate according to the function:

$$i = 3.0 + 2.0(\pi - \pi^*) \text{ with } \pi^* = 4.0,$$

and aggregate expenditure is the sum of:

$$\begin{aligned} C &= 200 + 0.75Y \\ I &= 85 - 2i \\ G &= 100 \\ X - IM &= 50 - 0.15Y - 3i \end{aligned}$$

(a) What is the equation for the ADπ curve?

(b) Plot the ADπ curve in a diagram, not necessarily to scale, that shows the horizontal intercept and slope of the curve.

Exercise 23.12 *Optional*: If careful research estimates potential output Y_P at 1,000, and the ADπ function is as derived in Exercise 23.11, what is the equilibrium inflation rate? What interest rate must the central bank set to defend its inflation target?

Chapter 24

Exchange rates, monetary policy, and fiscal policy

In this chapter we will explore:

In 1999 the Canadian economist Robert Mundell won the Nobel Prize in Economics for his work on the importance of exchange rate policy for the effectiveness of monetary and fiscal policies as tools to manage aggregate demand. Mundell showed for a small open economy like Canada, with a high degree of international capital mobility, that:

- With *flexible* exchange rates monetary policy is a powerful demand management tool, but fiscal policy is weak, and
- With *fixed* exchange rates monetary policy is ineffective as a demand management tool, but fiscal policy is strong.

The Bank of Canada's monetary policy framework, using interest rates, flexible exchange rates and an inflation rate target reflects these strong theoretical arguments. Furthermore, the difficulties European countries using the euro have experienced trying to cut budget deficits and stimulate income growth show the constraint a fixed exchange rate imposes on fiscal adjustments.

This chapter explains the foreign exchange market, flexible and fixed exchange rates and the reasons why different exchange rate policies affect the design and effectiveness of monetary and fiscal policy.

24.1 The balance of payments

The **balance of payments accounts** provide the background to supply and demand in the foreign exchange market. They record transactions between residents of one country and the rest of the world that involve payments in different national currencies. The balance of payments accounts provide the background to supply and demand in the foreign exchange market. They record transactions between residents of one country and the rest of the world that involve payments in different national currencies. For the Canadian Balance of Payments, entries in the accounts are converted to and entered in Canadian dollars. Transactions that give rise to an inflow of funds from other countries to Canada are entered as credits. Transactions that give rise to an outflow of funds from Canada to other countries are entered as payments.

Table 24.1 shows the actual Canadian balance of payments accounts in 2018. Entries in the 'Exports' column represent supplies of foreign currency coming to the Canadian foreign exchange market, converted to Canadian dollars. Entries in the 'Imports' column represent the demand for foreign currency in the Canadian foreign exchange market, converted to Canadian dollars. Row 5 in the table shows that after adjustment for statistical error and changes in official reserves, the foreign exchange market is in equilibrium – supply equals demand.

Balance of payments accounts: a record of trade and financial transactions between residents of one country and the rest of the world.

Table 24.1: The Canadian Balance of Payments, 2018
(Billions of Canadian $)

	Receipts (Exports)	Payments (Imports)	Balance
1. Current account			
Goods	585.3	607.2	−21.9
Services	120.8	145.8	−25.0
Investment income	130.6	136.8	−6.2
Transfers, etc.	13.9	19.3	−5.4
Balance			−58.5
2. Financial account			
Investment:	***Foreign in Canada***	***Canada in foreign***	
Direct investment:	58.8	68.3	−9.5
Portfolio investment	68.2	57.5	10.7
Other investment	81.2	37.1	46.0
Balance			**47.2**
3. Statistical discrepancy			**9.3**
4. Official reserves			−2.0
5. Balance of payments $[(1)+(2)+(3)-(4)]$			**0.0**

Source: Statistics Canada Tables 36-10-0014-01 and 36-10-0471-01

The **Current Account** of the balance of payments records international flows of goods, services, investment income and transfer payments. The merchandise trade is exports and imports of goods; things like cars and car parts, steel, wheat, and electronic equipment. Non-merchandise trade measures exports and imports of services like travel, banking and financial services, transportation, and tourism. The total of merchandise and non-merchandise trade is the trade balance defined as *net exports* in our earlier study of planned expenditure and aggregate demand.

> **Current Account**: a record of trade in goods, services, investment income and transfer payments.

However, the trade balance is not the same thing as the current account of the balance of payments. There are also flows of investment income in the form of interest payments, dividends and reinvested earnings, and transfer payments between countries as a result of government programs like foreign aid, and private receipts and payments.

But exports and imports of goods and services are the largest components of the current account. Trade in goods and services is based, in part on differences across countries in tastes, in the types of goods and services available, in economic structure, in national income levels and differences in the prices of domestic goods and services relative to foreign goods and services.

Three factors determine the prices of foreign goods and services relative to prices of domestic goods, namely:

1. The domestic price level (P_{Cdn}) for Canada;
2. The foreign currency price of imports (P_{US}) in the case of imports from the US; and
3. The **nominal exchange rate** (er): the domestic currency price of foreign currency.

International price competitiveness is measured by the **real exchange rate**, which combines these three factors. For example, the real exchange rate between Canada and the United States would be:

$$\text{Real exchange rate} = \frac{er \times P_{\text{US}}}{P_{\text{Cdn}}} \tag{24.1}$$

where the nominal exchange rate er is the Canadian dollar price of the US dollar and P_{US} and P_{Cdn} are general price levels as measured, for example, by GDP deflators or consumer price indexes. The real exchange rate measures the price of United States goods and services in Canadian dollars *relative to* Canadian goods and services in Canadian dollars.

The real exchange rate makes an important link in the transmission mechanism between the foreign exchange rate and the impact of domestic price levels and inflation rates on net exports, aggregate expenditure and aggregate demand.

> **Nominal exchange rate** (er): the domestic currency price of a unit of foreign currency.
>
> **Real exchange rate**: the relative price of goods and services from different countries measured in a common currency.

Consider the following example:

- The nominal exchange rate $er =$\$1.25Cdn/\$1US
- The GDP deflator for Canada is 121.3, on the base year 2002.
- The GDP deflator for the US is 110.4, on the base year 2002.

Then the real exchange rate, which gives the price of US goods in Canadian dollars relative to the price of Canadian goods in Canadian dollars, is:

$$\text{Real exchange rate} = \frac{er \times P_{\text{US}}}{P_{\text{Cdn}}} = \frac{1.25 \times 110.4}{121.3} = 1.138 \tag{24.2}$$

By this example, US goods and services are about 14 percent more expensive than Canadian goods and services, on average, when both are priced in Canadian dollars. At the same time, Canadian goods and services are less expensive than domestic goods in the US when both are priced in US dollars. These price differentials as are important to the flows of exports and imports of goods and services between countries.

The **Financial Account** of the balance of payments makes a link between interest rates and the supply and demand for foreign exchange and the exchange rate.

The financial account records the flows of funds for international purchases and sales of real and financial assets. Table 24.1 shows a net capital inflow of $47.2 billion in 2018. The payments by foreigners buying Canadian assets exceeded the payments made by Canadians buying foreign physical and financial assets. A financial account surplus was the result.

> **Financial Account**: the record of capital transfers and the purchases and sales of real and financial assets.

Receipts and payments in the financial account reflect sales and purchases of foreign assets. These flows have become increasingly important. Computers and electronic communications make it as easy for a Canadian resident to buy and sell stocks and bonds in the financial markets of New York or London as in Toronto. Moreover, controls on international capital flows have gradually been dismantled with globalization and financial integration.

The world's financial markets now have two crucial features. First, financial account restrictions on capital flows between advanced countries have been abolished. Funds can move freely from one country to another in search of the highest expected *rate of return*. Second, trillions of dollars are internationally footloose, capable of being switched between countries and currencies when assets with similar degrees of risk are expected to offer different rates of return in different countries and currencies.

This is the age of **perfect capital mobility** when small differences in expected returns trigger very large flows of funds from country to country. Indeed, the stock of international funds is now so huge that capital flows could swamp the typical current account flows from exports and imports.

> **Perfect capital mobility**: when very small differences in expected returns cause very large international flows of funds.

In international asset markets, capital gains or losses arise not merely from changes in the domestic price of an asset, but also from interest rate differentials and changes in exchange rates while temporarily holding foreign assets. Table 24.2 provides an example.

Table 24.2: Returns from lending $1,000 for a year

$1,000	Interest Rate (%)	Exchange rate ($Cdn/$US)		Final asset value	
Lent in		Initial	Final	$Cdn	$US
1. Canada	4.0			1,040.00	
2. United States	5.0	1.03	1.009	1,028.50	1,019.41

Suppose you can invest $1,000 Canadian for a year. Canadian one-year interest rates are 4 percent. In the United States one-year rates are 5 percent. The higher United States rates look attractive. If you keep your funds in Canadian dollars, Row 1 shows that you will have $1,040 at the end of the year. Can you do better by buying a United States asset?

Row 2 shows what happens if you convert $1,000Cdn into US dollars at an initial exchange rate of $1.03Cdn/$1US. You would have $970.87US to invest at the United States interest rate of 5 percent. You would get $1,019.41US. You would be ahead if the exchange rate remained constant. $1,019.41US is $1,050Cdn, a return of 5 percent, as you would expect.

Suppose, however, the exchange rate changes while your funds are out of the country. Let's say the Canadian dollar appreciates by 2 percent during the year, lowering the exchange rate to $1.009Cdn/$1US When converted back to Canadian dollars your $1,019.41US now buys $1,028.50 Cdn. You get 1 percent more interest income from holding the United States asset instead of the Canadian asset, but you suffer a capital loss of 2 percent by temporarily holding US dollars, whose value relative to Canadian dollars fell by 2 percent in that year.

In this example, you end up with about $1,028.50Cdn if you lend in US dollars. The Canadian dollar appreciated by more than 1 percent, the difference between Canadian and United States interest rates. As a result, the capital loss from the exchange rate while holding US dollars outweighed the gain on interest. This was the experience of many portfolios in 2008 as the Canadian dollar appreciated strongly and the exchange rate fell. The total return on lending in US dollars was lower than the return in Canadian dollars.

Conversely, if the Canadian dollar depreciated against the US dollar while you were holding your United States asset, your total return would be higher than the 1 percent interest rate differential. You would get a gain on the exchange rate when you converted back to Canadian dollars. This was the experience of portfolios holding assets denominated in US dollars as the Canadian dollar depreciated over the period from May 2008 to March 2009, raising the nominal exchange rate from $er = 0.9994$ to $er = 1.2645$. The exchange rate movement provided a 26.5% annual return, in terms of Canadian dollars, to portfolios holding US dollar assets. The depreciation of the Canadian dollar in 2014-15 had the same effect.

Equation 24.3 summarizes this important result. The total return on temporarily lending in a foreign currency is the interest paid on assets in that currency plus any capital gain (or minus any capital loss) arising from the depreciation (appreciation) of the domestic currency during the period.

$$\begin{array}{ccccc} \text{Return on holding} & = & \text{Interest rate on} & \pm & \text{\% increase/ \% decrease in} \\ \text{foreign asset} & & \text{foreign asset} & & \text{nominal exchange rate } (er) \end{array}$$

$$\text{Total return on foreign asset} = i_f + \%\Delta er \tag{24.3}$$

As a result, the net capital flow in the balance of payments depends positively on the differential between domestic and foreign nominal interest rates $(i - i_f)$. A rise in domestic rates relative to foreign rates would attract a flow of funds into the domestic financial market. A fall in domestic rates would push the flow toward foreign financial markets, assuming in both cases that the exchange rate is not expected to change in an offsetting direction.

Alternatively, assuming the interest rate differential is constant, the net capital flow depends negatively on the expected rate of depreciation of the domestic currency suggested by $[(er^e/er) - 1]$. An expectation that the domestic currency will depreciate (a rise in er^e) will increase the returns from holding foreign assets and lead to a net capital outflow. An expected appreciation would reduce expected returns on foreign assets with the opposite effect. These capital flows are important parts of the supply and demand for foreign exchange on the foreign exchange market.

The **change in official international reserves** in Table 24.1 records the increase or decrease in the Government of Canada's holdings of foreign currency balances. A government's holdings of foreign currencies are in its official international reserves account. These balances are like investments in foreign countries because they are the government's holdings of foreign assets. An increase in the official reserves is like a payment item in the financial account of the balance of payments. Because Canada maintains a flexible exchange rate, annual changes in international reserves are small. When it comes to discussing different exchange rate policies, countries that adopt fixed exchange rates often experience large changes in their foreign currency reserves in defense of the exchange rate they have set.

> **Change in official international reserves**: the change in the Government of Canada's foreign currency balances.

The **balance of payments** is the sum of the balances in current and financial accounts minus the change in the official international reserves account. In Table 24.1, this balance is shown as the sum of accounts $(1+2+3-4)$ namely $(-58.5+47.2+9.3-2.0) \approx 0$. If all items in the accounts were measured correctly, the balance would be zero. To recognize this, a statistical discrepancy adjustment is made, as shown in the table, to account for any errors in the measurement of other items.

> **Balance of payments**: the sum of the balances in current accounts and financial accounts, minus the change in the holdings of official reserves.

The record of the change in official reserves is always of equal magnitude to the sum of the balances on the current and financial accounts, if there is no statistical discrepancy in the measurements. As a result, the *balance of payments always balances*, but the state of the individual accounts underlying that overall balance need not be in balance. Indeed, changes in the foreign exchange rate in the foreign exchange market reconcile the different account balances to produce overall balance.

Net Domestic Savings, International Capital Flows and Policy Issues

The Balance of Payments must always balance, as the example in Table 24.1 shows. Current account surpluses or deficits must be offset by financial account deficits or surpluses. The National Accounts makes a link between incomes and expenditures and the current account in Balance of Payments that illustrates why and how *net domestic savings* and *net international capital flows* are related. That relationship often raises policy questions and debates about the causes of, and remedies for, current account deficits and government budget deficits.

To see this recall, from the National Accounts in Chapters 17 – 19, that GDP, the sum of *actual* expenditures, is:

$$GDP = C + I + G + X - IM$$

Then using *NNI*, the net investment income coming from transactions with other countries, as recorded in the Current Account of the Balance of Payments, *gross national product, GNP*, is:

$$GNP = C + I + G + X - IM + NII \tag{24.4}$$

Then the current account balance, *CAB*, net exports plus net investment income is:

$$CAB = X - IM + NII \tag{24.5}$$

and:

$$GNP = C + I + G + CAB \tag{24.6}$$

Furthermore, the *GNP* received by households is used for consumption or to pay taxes and to save:

$$GNP = C + T + S, \tag{24.7}$$

Now Balance of Payments balance comes from offsetting balances in the current account and the financial account. The current account balance, CAB, is the sum of net private saving and government saving. Then from (24.6) and (24.7) net private saving is the difference between private saving and private investment, namely: $(S-I)$. Net public saving is the difference between tax revenues and government expenditure, namely: $(T-G)$. Then the current account balance, CAB, is:

$$CAB = (S-I) + (T-G) \tag{24.8}$$

This equation says that changes in the current account balance are a result of one of:

a) A change in private sector saving or investment spending, or
b) A change in the government budget balance, or
c) A change in net exports $(X - IM)$

without an offsetting change in the other balances.

The effects of changes in economic conditions are more interesting. For example, a rise in domestic investment that reduces net private saving will lead to a fall in the current account balance. Alternatively, a cut in government expenditures that increases the government's budget balance will lead to a rise in the current account balance. This link between the government balance and the current account balance is the basis of the 'twin deficits' argument that current account deficits are caused by government budget deficits and should be addressed by fiscal austerity rather than import restrictions. But an increase in private sector investment that increases business capacity or residential housing stock reduces net private saving and would also reduce the current account balance. The important policy questions are whether the underlying cause of the change in the current account balance would positively affect national economic conditions and activity.

However, regardless of the causes of changes in the current account balance, international capital flows in the financial account balance will change to maintain overall balance in the Balance of Payments.

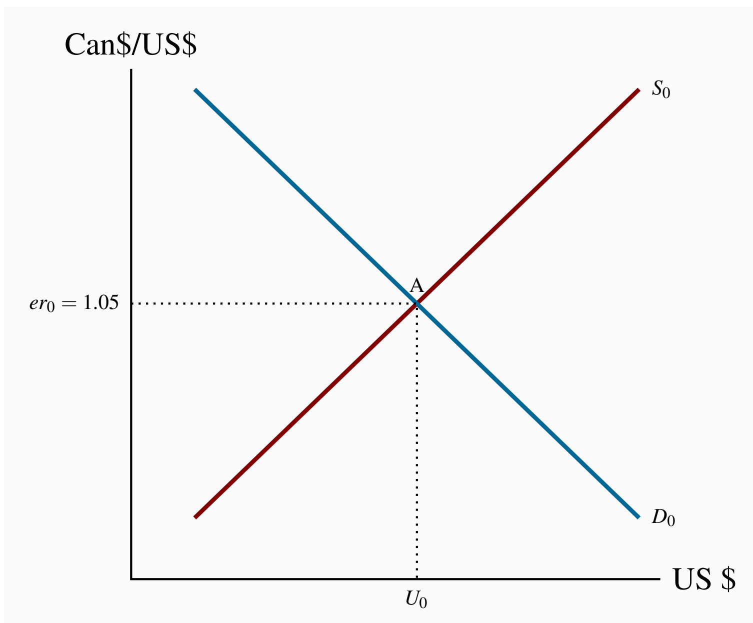
24.2 The foreign exchange market

The foreign exchange market is the market in which the currencies of different countries are bought and sold and the prices of currencies, the foreign exchange rates, are established. Consider the market for US dollars as the foreign currency. The sources of supply and demand for foreign exchange are shown by the balance of payments in Table 24.1. Exports of goods, services, and financial assets generate a supply of foreign currencies that are sold in the foreign exchange market for

Canadian dollars. Imports of goods, services, and securities must be paid for in foreign currencies. The demand for foreign exchange is derived from this demand for imports. Without intervention by governments, demand and supply determine the exchange rate, as, for example, $er_0 = 1.05$ in Figure 24.1.

Different economic and financial conditions would change supply and demand for foreign exchange and result in a different exchange rate. In May of 2019, for example, the Canadian dollar price of the US dollar was \$1.33. By contrast when energy and commodity prices were high in 2010 the price of a US dollar was \$1.0299 Canadian dollars.

Figure 24.1: The foreign exchange market



The supply of US dollars on the foreign exchange market comes from US purchases of Canadian goods, services and assets. The demand for US dollars comes from Canadian demand for imports of goods, services and assets. Supply and demand determine the equilibrium exchange rate er_0.

The exchange rate $er_0 = \$1.05$ is what it costs in Canadian dollars to buy each US dollar you want for your winter reading week break in Florida. Alternatively, if you as an exporter of lumber to the US market receive \$1,000US for every 1000 board feet of 2x4 lumber you sell to US builders, your Canadian dollar revenue to cover your Canadian costs is \$1,050.

If the price of goods in US markets is constant, a lower exchange rate er_1; say, $er_1 = 1.01$, *other things constant*, by lowering the price and raising the quantity of Canadian imports, must raise the demand for US dollars. In Figure 24.1, the demand curve for US dollars D_0 *slopes downwards*. More US dollars are demanded to buy more imports at a lower Cdn\$/US\$ exchange rate.

The US dollars supplied on the foreign exchange market are the receipts from the export of goods, services, and securities to US residents. From the discussion of the current and financial accounts of the balance of payments, exports of goods and services depend on foreign income, the relative prices of domestic and foreign goods and services, and the exchange rate. Net exports of securities depend on the difference between domestic and foreign interest rates, for given expectations of the future exchange rate.

In Figure 24.1, the supply curve shows the quantities of US dollars that would come to the market at different exchange rates, er, all other things constant. It slopes upward because a higher Cdn$/US$ exchange rate ($er > \$1.05$) lowers the prices of Canadian goods and services to US buyers. As a result, United States residents buy more Canadian exports and total export receipts rise. The quantity of US dollars coming onto the foreign exchange market increases as we move up the supply curve. The supply curve has a positive slope.

Figure 24.1 assumes that the demand for Canadian exports and the Canadian demand for imports are price elastic. *Price increases caused by changes in the exchange rate reduce total revenue. Price reductions caused by changes in the exchange rate increase total revenue*. These conditions give the slopes shown for the supply and demand curves.

At the equilibrium exchange rate er_0, the quantities of US dollars supplied and demanded are equal. In terms of the balance of payments, the balance is zero. Receipts equal payments.

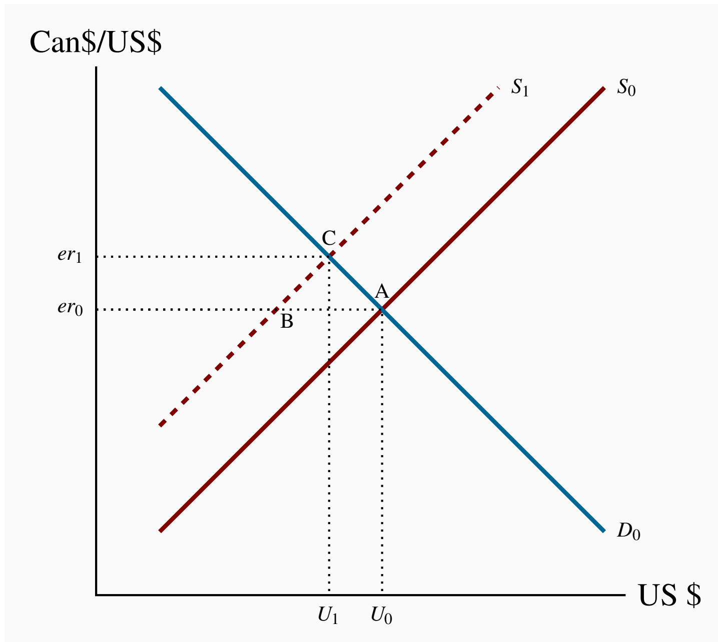
In practice, as seen in Table 24.1, the balance of payments is recorded in domestic currency. However, the equality of receipts and payments still holds. Both are easily converted to Canadian dollars by multiplying the US dollar amounts by the exchange rate. In terms of Figure 24.1, multiplying U_0 on the horizontal axis by the exchange rate er_0 on the vertical axis gives the area of the rectangle er_0AU_0O, the Canadian dollar value of total receipts or total payments recorded in the balance of payments.

What would change the equilibrium in Figure 24.1? A change in any of the factors held constant in order to draw the supply and demand curves will shift one or the other or both curves. We see this in both the net export and capital flow functions. A rise in United States income would increase US imports from Canada and shift the supply of foreign exchange to the right. As discussed above and in Chapter 21, a change in interest rates in Canada or the United States would change the trade in financial assets and affect both the supply curve and the demand curve. In short, a change in any market condition *other than the exchange rate er* will change supply and demand conditions in the market. The exchange rate will then change to a new equilibrium. Figure 24.2 and 24.3 provide examples.

The effect of a recession in the US economy on the exchange rate

The demand and supply curves in the foreign exchange market of Figure 24.2 are drawn on the assumption that tastes, incomes, prices of goods and services, interest rates, and expectations of future exchange rates are constant. The flows of payments and receipts under these conditions result in the equilibrium exchange rate er_0. This would be a Canadian dollar price for the US dollar of, for example, \$1.35. From the United States perspective, a Canadian dollar costs a United States resident about \$0.74US.

Figure 24.2: The effect of a recession in the US on the exchange rate



A recession in the US reduces Canadian exports to the US and reduces the supply of US\$ on the foreign market. S shifts from S_0 to S_1. The exchange rate rises as the Can\$ depreciates.

Now suppose, as occurred in early 2008 and 2009, a recession in the United States lowers United States real income. United States imports fall, based on the US marginal propensity to import. United States imports are Canadian exports, and the US dollar receipts of Canadian exporters are reduced. The recession and difficult household financial conditions reduce US residential construction and Canadian lumber exports decline. Recession also reduces travel by US residents, and the Canadian tourism industry suffers a decline in bookings and receipts. If expenditures on new cars in the United States are reduced, Canadian auto industry sales to the United States market are reduced. In the balance of payments, the balance on trade in goods and services falls, and in the foreign exchange market *the supply of US dollars on the market is reduced*.

The supply curve in Figure 24.2 shifts leftward to S_1. At the initial exchange rate er_0, the demand for US dollars exceeds the supply, putting upward pressure on the exchange rate. In terms of the balance of payments, the excess demand for US dollars represents a balance of payments deficit. In the example shown here, the Canadian dollar depreciates and the exchange rate rises to restore equilibrium in the foreign exchange market and the balance of payments. A higher price for the US dollar reduces Canadian imports and increases the Canadian dollar receipts of Canadian exporters. The effects of the US recession on the Canadian balance of payments are offset by the exchange rate change.

The rise in the exchange rate in this case is a **depreciation** of the Canadian dollar and a corresponding **appreciation** of the US dollar.

> **Currency depreciation**: a fall in external value of the domestic currency that raises domestic currency price of foreign currency.
>
> **Currency appreciation**: a rise in external value of the domestic currency that lowers the domestic currency price of foreign currency.

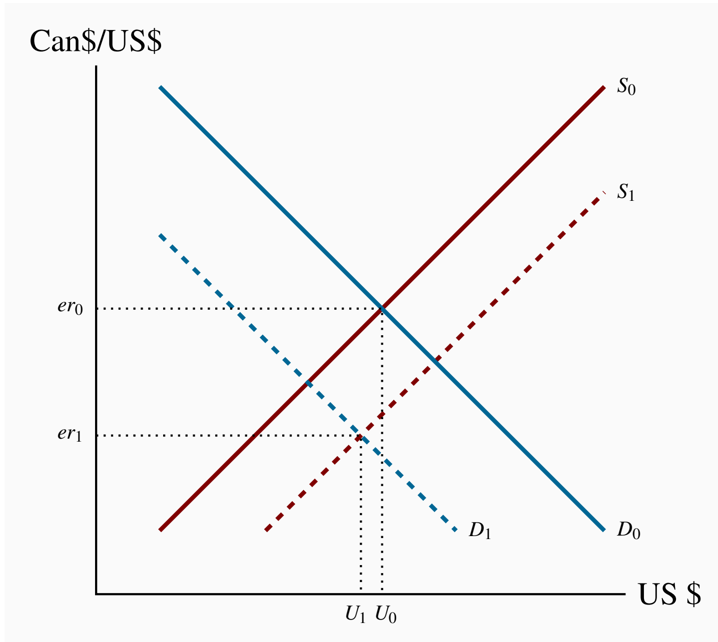
Before the recession of 2009, Canadian experience was the opposite of this example. High GDP growth rates in the US and Asia created very strong international demand for Canadian commodities and crude oil at high international prices. Strong oil and commodity exports increased the supply of foreign currencies on the Canadian foreign exchange market. The Canadian dollar *appreciated* strongly, with the exchange rate falling from $1.57Cdn/$1US in 2002 to an average of $0.9994 Cdn/$1US in May of 2008, a fall of about 60 percent over the six year period. Exchange rate changes led to a restructuring of both exports and imports to maintain equilibrium in the balance of payments as Canada's international trade changed dramatically.

The effect of a fall in foreign interest rates on the foreign exchange rate

In the previous example a change in foreign income and the supply of foreign exchange disrupted the equilibrium in the foreign exchange market and changed the exchange rate. As an alternative to that example, consider the effects of a cut in foreign interest rates.

In Figure 24.3 the foreign exchange market is in equilibrium, initially at an exchange rate er_0. A fall in foreign interest rates, other things constant, disrupts this equilibrium. Now lower foreign rates make domestic (Canadian) bonds more attractive to foreign portfolios than they were previously. The demand for domestic bonds rises. The supply of US dollars coming on the market to pay for these bond exports increases, shifting S_0 to S_1 in the diagram. At the same time, the attractiveness of foreign bonds for domestic portfolios is reduced, reducing the demand for US dollars to pay for them. The demand curve shifts from D_0 to D_1. The equilibrium exchange rate falls to er_1.

Figure 24.3: The effect of a cut in foreign interest rates



A cut in foreign interest rates shifts portfolios toward the Canadian bond market and away from foreign markets. The supply of foreign currency increases from the increased export of bonds at the same time as the demand for foreign currency to buy foreign bonds falls. The domestic currency appreciates and the exchange rate falls.

In balance of payments terms the net capital inflow, the balance on financial account, is increased. The change in the exchange rate causes an offsetting change in the current account that restores equilibrium in the foreign exchange market and the balance of payments. A lower nominal exchange rate lowers the real exchange rate. Imports are now cheaper and exports, priced internationally in US dollars, are less profitable. Export receipts are reduced. As we saw in Chapter 21, changes in the exchange rate are one channel by which changes in financial conditions impact AD and equilibrium GDP.

Figures 24.2 and 24.3 provide two examples of adjustments in the foreign exchange rate. The underlying assumption is that *exchange rates are flexible* and allow the market to adjust freely and quickly to changing circumstances.

However, there are alternative exchange rate arrangements. To understand how the foreign exchange rate operates in different countries, we need to consider the different *exchange rate policies* governments can adopt. These result in different foreign exchange rate regimes, different ways that the balance of payments adjusts to change, and different roles for monetary and fiscal policies.

24.3 Flexible exchange rates and fixed exchange rates

To grasp the basics of **exchange rate regimes**, we focus on two extreme forms that have been adopted to handle international transactions in the world economy: *flexible exchange rates and fixed exchange rates*.

> **Exchange rate regime**: the policy choice that determines how foreign exchange markets operate.

Floating or flexible rates

In a **floating or flexible exchange rate** regime, the exchange rate is allowed to find its equilibrium level on the foreign exchange market *without central bank intervention*.

> **Flexible exchange rates**: Supply and demand in the foreign exchange market determine the equilibrium exchange rate without central bank intervention.

Figures 24.2 and 24.3 showed the exchange rates that would result if rates adjusted flexibly and freely in response to changes in demand and supply. The central bank *did not intervene* to fix or adjust the rate. The rise in the demand for US dollars would result in a rise in the exchange rate to clear the foreign exchange market and maintain the balance of payments. Alternatively, the fall in demand would result in a fall in the exchange rate. The Bank of Canada would not intervene in either case. The holdings of official foreign exchange reserves and the domestic money supply would not be affected by foreign exchange market adjustments.

The alternative is a *fixed exchange rate* as explained below. In this regime, the central bank sets an official exchange rate and intervenes in the foreign exchange market to offset the effects of fluctuations in supply and demand and maintain a constant exchange rate. In Canada in the 1960s the exchange rate was fixed by policy at \$1US=\$1.075 Cdn (\$1Cdn≈\$0.925US) and the Bank of Canada intervened in the foreign exchange market to maintain that rate.

How do countries choose between fixed and floating exchange rates? Obviously, there is not one answer for all countries or we would not see different exchange rate regimes today. With flexible rates, the foreign exchange market sets the exchange rate, and monetary policy is available to pursue other targets. On the other hand, fixed exchange rates require central bank intervention. *Monetary policy is aimed at the exchange rate*.

The importance a country attaches to an *independent monetary policy* is one very important factor in the choice of an exchange rate regime. Another is the size and volatility of the international trade sector of the economy. A flexible exchange rate provides some *automatic adjustment and stabilization* in times of change in net exports or net capital flows.

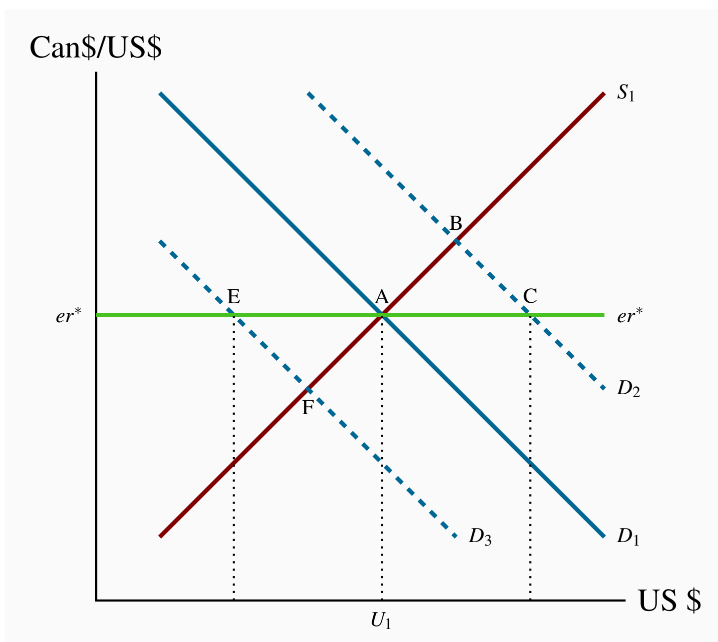
Fixed exchange rates

In a **fixed exchange rate** regime, the government intervenes actively through the central bank to maintain convertibility of their currency into other currencies at a fixed exchange rate. A currency is **convertible** if the central bank will buy or sell as much of the foreign currency as people wish to trade at a fixed exchange rate.

> **Fixed exchange rate**: an exchange rate set by government policy that does not change as a result of changes in market conditions.
>
> **Convertible currency**: a national currency that can be freely exchanged for a different national currency at the prevailing exchange rate.

In Figure 24.4, suppose the exchange rate is *fixed at* er^*. There would be a free market equilibrium at A if the supply curve for US dollars is S_1 and the demand curve for US dollars is D_1. The central bank does not need to buy or sell US dollars. The market is in equilibrium and clears by itself at the fixed rate.

Figure 24.4: Central bank intervention to fix the exchange rate



With exchange rate fixed at er^* a shift in demand for US$ to D_2 creates excess demand AC. The central bank intervenes, supplying AC US$ from official reserve holdings in exchange for Can$. To maintain er^* if demand shifted to D_3 would create the opposite condition and central bank would have to buy US$.

Suppose demand for US dollars shifts from D_1 to D_2. Canadians want to spend more time in

Florida to escape the long, cold Canadian winter. They need more US dollars to finance their expenditures in the United States. The free market equilibrium would be at B, and the exchange rate would rise if the Bank of Canada took no action.

However, with the exchange rate fixed by policy at er^* there is an excess demand for US dollars equal to AC. To peg[1] the exchange rate, the Bank of Canada *sells* US dollars from the **official exchange reserves** in the amount AC. The supply of US dollars on the market is then the "market" supply represented by S_1 plus the amount AC supplied by the Bank of Canada. The payment the Bank receives in Canadian dollars is the amount ($er^* \times$AC), which *reduces the monetary base* by that amount, just like an open market sale of government bonds. The lower monetary base pushes domestic interest rates up and attracts a larger net capital inflow. Higher interest rates also reduce domestic expenditure and the demand for imports and for foreign exchange. *The exchange rate target drives the Bank's monetary policy*, which in turn changes both international capital flows and domestic income and expenditure.

> **Official exchange reserves**: government foreign currency holdings managed by the central bank.

What if the demand for US dollars falls to D_3? The market equilibrium would be at F. At the exchange rate at er^* there is an excess supply of US dollars EA. The Bank of Canada would have to *buy* EA US dollars, reducing the supply of US dollars on the market to meet the "unofficial" demand. The Bank of Canada's purchase would be added to foreign exchange reserves. The Bank would pay for these US dollars by *creating more monetary base*, as in the case of an open market purchase of government securities. In either case, maintaining a fixed exchange rate requires **central bank intervention** in the foreign currency market. The central bank's monetary policy is expansionary because it is committed to the exchange rate target.

> **Central bank intervention**: purchases or sales of foreign currency intended to manage the exchange rate.

When the demand schedule is D_2, foreign exchange reserves are running down. When the demand schedule is D_3, foreign exchange reserves are increasing. If the demand for US dollars fluctuates between D_2 and D_3, the Bank of Canada can sustain and stabilize the exchange rate in the long run.

However, if the demand for US dollars is, on average, D_2, the foreign exchange reserves are steadily declining to support the exchange rate er^*, and the monetary base is falling as well. In this case, the Canadian dollar is overvalued at er^*; or, in other words, er^* is too low a price for the US dollar. A higher er is required for long-run equilibrium in the foreign exchange market and the balance of payments. As reserves start to run out, the government may try to borrow foreign exchange reserves from other countries and the International Monetary Fund (IMF), an international body that exists primarily to lend to countries in short-term difficulties.

[1] Typically, a country will "peg" its currency to a major currency such as the US dollar, or to a basket of currencies.

At best, this is only a temporary solution. Unless the demand for US dollars decreases, or the supply increases in the longer term, it is necessary to **devalue** the Canadian dollar. If a fixed exchange rate is to be maintained, the official rate must be reset at a higher domestic currency price for foreign currency.

> **Devaluation (revaluation)**: a reduction (increase) in the international value of the domestic currency.

For many years frequent media and political discussions of the persistent rise in China's foreign exchange holdings provide a good example of the defense of an undervalued currency. With the yuan at its current fixed rate relative to US dollars and other currencies, China has a large current account surplus that is not offset by a financial account deficit. Balance of payments equilibrium requires ongoing intervention by the Chinese central bank to buy foreign exchange and add to official reserve holdings. Buying foreign exchange adds to the monetary base and money supply, raising concerns about inflation. The Bank has responded in part with a small revaluation of the yuan and in part with an increase in the reserve requirements for Chinese banks. Neither of these adjustments has been sufficient to change the situation fundamentally and growth in official foreign exchange reserves continues.

In Europe the euro currency system effectively fixes exchange rates among member countries. Individual member countries do not have national monetary policies. Monetary policy is set by the European Central Bank. In the years following the 'great recession' this has been a source of controversy because economic and fiscal conditions have differed significantly among countries. Countries trying to adjust fiscal deficits and national public debt crises have been forced into fiscal austerity without offsetting monetary policy support. In many cases the results have been deep and prolonged recessions without solving their debt problems. Greece is the poster child.

Of course, it is not necessary to adopt the extreme regimes of pure or clean floating on the one hand and perfectly fixed exchange rates on the other hand. *Dirty or managed floating* is used to offset large and rapid shifts in supply or demand schedules in the short run. The intent is to smooth the adjustment as the exchange rate is gradually allowed to find its equilibrium level in response to longer-term changes.

24.4 Monetary and fiscal policy under flexible exchange rates

In a closed economy with slow wage and price adjustments, monetary and fiscal policies are both important tools for aggregate demand management in the short run. Things are different in open economies with high international capital mobility. *With flexible exchange rates monetary policy is powerful* for changing AD. It works through both interest rate and exchange rate linkages in the transmission mechanism, not just the interest rate linkages of the closed economy. By contrast, *the effects of fiscal policy on aggregate demand are reduced*. In the absence of supporting monetary policy, fiscal expansions crowd out private sector expenditures through both interest rate and exchange rate linkages, leaving AD unchanged.

Fixed exchange rates have the opposite implications for policy effectiveness as an AD management tool. *The effects of fiscal policy are enhanced by induced changes in monetary conditions required to maintain the fixed exchange rate. But monetary policy alone, committed to fixing the exchange rate, is almost powerless to change AD*. As a result, the first important step in the design of macroeconomic policy in the open economy is the choice of an exchange rate regime.

Monetary policy

With flexible exchange rates, monetary policy causes changes in both interest rates and exchange rates. Net international capital flows link exchange rates and changes in domestic interest rates when exchange rates are flexible. Given the exchange rate expected in the long run, higher interest rates in the short to medium run cause a capital inflow, an increased supply of foreign exchange on the foreign exchange market, which lowers the exchange rate, *er*. Conversely, lower domestic interest rates relative to international rates cause a rise in the exchange rate, *er*.

The 'transmission mechanism' introduced in Chapter 21 shows how changes in monetary policy, with flexible exchange rates change aggregate demand and equilibrium income.

$$\Delta i_{onr} \rightarrow \begin{pmatrix} \Delta C + \Delta I \\ \Delta er \rightarrow \Delta NX \end{pmatrix} \rightarrow \Delta AE \times \text{ multiplier } \rightarrow \Delta AD \rightarrow \Delta Y$$

As a result, in an open economy with flexible exchange rates, monetary policy affects aggregate demand not just through the effects of interest rates on consumption and investment. With linkages through both domestic and international components of expenditure, monetary policy is more powerful under flexible exchange rates than in a closed economy.

Canada and a number of other countries conduct monetary policy in terms of a target for the domestic inflation rate. A flexible exchange rate policy is essential for the monetary policy independence and power required to pursue that target. That is why the Bank of Canada defines Canada's flexible exchange rate as a key component of Canada's monetary policy framework. The other key component is the Bank's inflation control target.

Fiscal policy

With flexible exchange rates, but without monetary policy accommodation or support, the effect of interest rate changes on exchange rates and competitiveness undermines the power of fiscal policy to manage aggregate demand.

Suppose the government undertakes a *fiscal expansion*, raising government expenditures or lowering taxes or some combination of the two: Aggregate demand increases. When monetary policy targets an inflation rate using either an interest rate rule or a money supply rule, the expansion in AD caused by fiscal policy changes the economic fundamentals on which the central bank's policy had been set and induces the bank to *raise interest rates*. The higher interest rates cause a net capital inflow and an increased supply of foreign exchange on the foreign exchange market, and the nominal exchange rate falls. A fall in the nominal foreign exchange rate *lowers the real exchange rate*. International price competitiveness (as measured by the real exchange rate) is reduced and net exports fall, offsetting the expansionary effects of the change in fiscal policy. *With flexible exchange rates monetary policy targeted to the inflation rate dominates fiscal policy* as a tool for AD management. The sequence of crowding out would be:

$$\uparrow G \text{ or } \downarrow T \rightarrow \uparrow AE \rightarrow \frac{\uparrow Y}{Y_P} \rightarrow \uparrow \pi \rightarrow \uparrow i_{onr} \rightarrow \downarrow er \rightarrow \downarrow NX \rightarrow \downarrow AE \rightarrow \frac{\downarrow Y}{Y_P}$$

As discussed previously, Canadian experience provides an example. In the 2005-2007 period, the federal government provided fiscal stimulus through tax cuts and expenditure increases. The primary structural budget balance fell from an average 3.4 percent of potential GDP for 2002 to 2004 to 2.6 percent for 2005 to 2007. At the same time the Bank of Canada's estimates showed the economy operating with a small but persistent inflationary gap. The inflation rate was in the upper level of the Bank's target range. The Bank responded to strong current and expected demand (coming from both the government and private sector) by raising its overnight rate in steps, from 2.5 percent in late-2005 to 4.5 percent by mid-2007, to defend its inflation target. The inflationary gap and inflation were contained as higher interest rates and lower exchange rates limited the growth of aggregate demand, including that coming from fiscal stimulus.

However, if control or reduction of the debt ratio is the prime target of fiscal policy, the flexible exchange rate is helpful. If the government raises tax rates or cuts expenditures to raise its structural budget balance and reduce the debt ratio, lower settings for the central bank's interest rate and a rising exchange rate provide some offsetting "crowding in" through both domestic expenditure and net exports. The limited aggregate demand effects of fiscal policy under flexible exchange rates facilitate control of the government's budget balance and debt ratio.

This analysis of the policy implications of flexible rate regimes leads to a clear recommendation for policy co-ordination. Flexible exchange rates provide the framework for effective monetary policy focused on a medium term inflation target. The exchange rate regime enhances the power of monetary policy to moderate business cycle fluctuations and the output gaps they create. Stabilizing the economy at or close to potential output avoids the cumulative inflationary or recessionary pressures that would push inflation rates away from the monetary policy target. Monetary policy is then the aggregate demand management tool.

Fiscal policy is not an effective AD management tool when exchange rates are flexible. Its impacts on aggregate demand are limited by crowding out and dominated by monetary policy. However, this does enhance the power of fiscal policy to pursue deficit control and debt ratio control. The effects of fiscal restraint aimed at improved public finances are moderated by a monetary policy focused on an inflation target in a flexible exchange rate regime.

The Canadian experience with economic policy and performance provides an excellent example of this sort of coordinated policy. Starting in 1995 the federal government introduced a policy of strong fiscal adjustment through restraint aimed at reducing the public debt-to-GDP ratio. The structural primary budget balance was increased through expenditure cuts and tax increases. At the same time, monetary policy was aimed at maintaining inflation within the 1-3 percent target band, which required monetary stimulus. The nominal and real overnight interest rates were reduced. Economic growth was constrained by the fiscal austerity but still sufficient to eliminate the recessionary gap by the end of the decade. The support of domestic monetary policy, a significant depreciation of the Canadian dollar, strong growth in the US economy and strong export growth were keys to this successful fiscal adjustment. The coordination of fiscal restraint and monetary stimulus moved the economy to potential output with stable inflation and a falling ratio of public debt to GDP.

Policy responses to the recession of 2009 provide a further example, both domestic and international. Monetary policy was the first line response, with central banks in most industrial countries lowering their interest rate to or close to the **zero lower bound**. Some countries, like the US, then went further to provide quantitative and credit easing through general and selective open market operations. Fiscal stimulus added to these highly accommodative monetary conditions. Central bank commitments, like those in both the US and Canada, to maintain policy rates at their minimum for periods as long as a year or more eliminated concerns about fiscal crowding out.

Coordinating the focus of both monetary policy and fiscal policy was designed to stimulate aggregate demand and restore growth in real GDP. In a time of deep recession, high indebtedness, and high uncertainty even very low interest rates won't induce households and business to take on more debt to build more houses or factories. There is already an excess supply of productive capacity and housing. Monetary conditions can support an expansion in expenditure but cannot trigger it.

Fiscal policy, by contrast, can add directly to expenditure and aggregate demand, especially expenditure on infrastructure, education, research, and similar public investments. Tax cuts are likely to have smaller expenditure effects if only because the recipients have marginal propensities to spend that are less than one. Nonetheless, there is an important debate about whether expenditure increases or tax cuts should be used for fiscal stimulus, and which will have the larger and more desirable effect.

With this policy coordination there is no cause for concern about crowding out. Central banks were not concerned about the effects of increased aggregate demand on inflation rates and their inflation targets. Quite the opposite, like the Bank of Canada they hoped to raise inflation to their target. Fiscal expansion will not induce higher interest rates or lower exchange rates.

24.5 Monetary and fiscal policy under fixed exchange rates

Monetary policy

If a country adopts a fixed exchange rate policy, *the exchange rate is the target of monetary policy*. Monetary policy cannot pursue an inflation target or an output target at the same time as it pursues an exchange rate target. Nor can it set either interest rates or money supply growth rates independently.

With a fixed exchange rate, interest rates must be set as needed to maintain the exchange rate when capital mobility is high. Indeed, the higher international capital mobility is, the less is the scope for independent monetary policy. This is what we mean when we say fixed exchange rates eliminate monetary policy sovereignty. The central bank cannot follow an independent monetary policy.

Fiscal policy with fixed exchange rates

A fixed exchange rate and perfect capital mobility undermine the scope for monetary policy, but maintain the effectiveness of fiscal policy.

In a closed economy, in the short run, fiscal expansion raises output. Under a monetary policy rule with the interest rate as the policy instrument, *as long as output is less than potential output*, the central bank supports the increase in output by maintaining interest rates and increasing the money supply as output expands. However, at outputs equal to or greater than potential output, central banks raise interest rates to crowd out the effect of fiscal expansion.

In an open economy with fixed exchange rates, monetary policy adjusts passively to keep the interest rate fixed in order to defend the exchange rate. Interest rates do not change to support fiscal policy or moderate the effect of fiscal policy. Hence, any fall in domestic demand can be offset by a fiscal expansion to help restore potential output. If the change in domestic demand is the only reason that the current account balance departed from equilibrium, this fiscal expansion will also restore the current account balance.

Fiscal policy is an important stabilization policy under fixed exchange rates. It helps to compensate for the fact that monetary policy can no longer be used. Automatic fiscal stabilizers play this role. Discretionary changes in government spending or taxes are useful only if fiscal policy can react quickly to temporary shocks. In some political systems, such as in Canada, this is feasible. In others, such as in the United States, where Congress and the President may be from different parties and budget decisions are more protracted, rapid changes in fiscal policy are more difficult.

In times of prolonged recession, discretionary fiscal policy can contribute importantly to a return to potential output, *provided it is not constrained by high public debt ratios*. With interest rates tied to the exchange rate, financing a fiscal expansion does not push rates up to crowd out private sector expenditure; nor does the recovery of the economy result in rising rates. Indeed, fiscal policy is the only effective domestic demand management tool available.

Unfortunately, high public debt ratios and concerns about the default risk of sovereign debt cause problems for fiscal policy. This was the situation recently in Europe. Fiscal expansion is impossible if financial markets are unwilling to buy more sovereign debt from economies in recession that already have high debt ratios. On the other hand, fiscal austerity to control deficits and debt ratios makes recessions worse and may even raise already high debt ratios as GDP and government net revenue fall. The euro fixes exchange rates within Europe and precludes stimulus from currency depreciation. Neither domestic fiscal nor monetary policy offers a solution.

Things are further complicated because many European countries have similar economic and financial difficulties. Canadian success with fiscal austerity and adjustment in the 1990s came from a very different economic environment. Monetary policy and exchange rate depreciation provided stimulus. Strong economic growth in major trade partners provided further stimulus through export growth. These conditions are clearly not met in Europe and coordinated fiscal austerity to address sovereign debt issues has been described as an 'economic suicide pact'.

Next

This chapter extended the discussion of short-run macroeconomic performance and policy by covering in more detail the importance of international trade, capital flows, and exchange rates for the design, coordination and effectiveness of monetary and fiscal policies. The next chapter introduces the theory of economic growth that explains the long-run growth in potential output based on growth in the labour force, growth in the capital stock and changes in productivity based on advances in technology.

Key Concepts

The **balance of payments** records transactions between residents of one country and the rest of the world. The **current account** shows the trade balance plus net international transfer payments, and income earned on holdings of foreign assets. The **financial account** shows net purchases and sales of foreign assets. The balance of payments is the sum of the current and financial account balances.

The trade in goods and services recorded in the current account is **net exports**, based on tastes, incomes, and the real exchange rate, which measures the price of foreign goods and services relative to the price of domestic goods and services.

The **trade in financial** assets recorded in the financial account is based on the total return expected from holding foreign rather than domestic assets.

The total return on holdings of foreign assets depends on the interest rate differential between countries and the change in the exchange rate during the period in which assets are held. **Perfect international capital mobility** means that an enormous quantity of funds shifts between currencies when the perceived rate of return differs across currencies.

The **foreign exchange market** is the market in which currencies of different countries are bought and sold and foreign exchange rates are established. The **exchange rate** is the price at which one currency trades for another.

The **demand for foreign currency** on the foreign exchange market arises from imports of goods and services and purchases of foreign assets. The **supply of foreign currency** on the foreign exchange market arises from exports of goods and services and sales of domestic assets to foreigners.

Under a **fixed exchange rate regime**, a balance of payments surplus or deficit must be matched by an offsetting quantity of **official financing**. The **central bank** intervenes in the foreign exchange market.

Under **floating or flexible exchange rates**, supply and demand in the foreign exchange market change the exchange rate as necessary for a current account balance that offsets a capital account balance. As a result, the balance of payments is zero and *no official intervention* is involved.

The **choice between fixed and floating exchange rate regimes** reflects a country's assessment of *the importance of an independent monetary policy*, the volatility of exports and imports, and the financial discipline that may come with fixed rates.

Flexible exchange rates increase the *effectiveness of monetary policy* as a tool to manage aggregate demand. *The effectiveness of fiscal policy* for demand management is reduced, but pursuit of deficit and debt ratio control may be enhanced.

Monetary policy sovereignty is lost when **fixed exchange rates** are adopted. Monetary policy cannot effectively pursue domestic inflation or output targets. However, the *effectiveness of fiscal policy as a demand management tool is enhanced*.

Exercises for Chapter 24

Exercise 24.1 Consider a country with a fixed exchange rate that has a current account surplus of $20 billion, but a financial account deficit of $18 billion.

(a) Is its balance of payments in deficit or surplus? Why?
(b) What change in official exchange reserves would you see? Why?
(c) Is the central bank buying or selling foreign currency?
(d) What effect does the central bank's foreign currency purchase or sale have on the monetary base? Explain why.

Exercise 24.2 Assume the initial exchange rate is $1.20Cdn for $1.00US After 10 years, the United States price level has risen from 100 to 200, and the Canadian price level has risen from 100 to 175. What was the inflation rate in each country? What nominal exchange rate would preserve the initial real exchange rate? Which country's currency depreciated?

Exercise 24.3 Suppose portfolio managers shift $100 million in assets under their control out of Canadian government securities and into United States government securities. What change would this portfolio shift make in the Canadian balance of payments?

Exercise 24.4 What is the expected rate of appreciation of the US dollar if interest rate parity prevails and Canadian nominal interest rates are 1 percent higher than United States interest rates?

Exercise 24.5 Suppose natural gas and crude oil prices were to drop sharply and expectations were they would remain low. Use a foreign exchange market diagram to show the effect on the Canadian/US dollar exchange rate?

Exercise 24.6 Using a diagram to illustrate:

(a) The demand for foreign exchange and the demand curve for foreign exchange.
(b) The supply of foreign exchange and the supply curve for foreign exchange.
(c) The equilibrium exchange rate.

Exercise 24.7 Use a foreign exchange market diagram with a flexible or floating exchange rate to show:

(a) How a decline in exports would affect the foreign exchange rate.

(b) How exports and imports would change to give balance of payments equilibrium at the new equilibrium exchange rate.

(c) The effects, if any, on the holdings of official reserves.

Exercise 24.8 Use a foreign exchange market diagram to show:

(a) Equilibrium with a fixed exchange rate.

(b) The effect of a decline in exports on conditions in the foreign exchange market when the exchange rate is fixed.

(c) The amount of the purchase or sale of foreign exchange reserves required if the central bank defends the fixed exchange rate.

(d) The effects of a change in the holdings of official reserves and the monetary base as a result of the defense of the fixed exchange rate.

Exercise 24.9 Use AD/AS and foreign exchange market diagrams to show why monetary policy is powerful and fiscal policy is weak when a country has a flexible exchange rate regime.

Exercise 24.10 Use AD/AS and foreign exchange market diagrams to show why the choice of a fixed exchange rate makes fiscal policy a more powerful tool for demand management. What happens to the domestic money supply when a government austerity program cuts its expenditures on goods and services and raises taxes?

Chapter 25

Economic growth

In this chapter we will explore:

Economic growth is one of the most historic, challenging and important topics in macroeconomics. Standards of living within a country are measured by real GDP *per capita*. Growth in real GDP is growth in potential output, Y_P, and growth in real GDP *per capita* is a measure of change in the standard of living. Over time, growth in Y_P that exceeds population growth raises per capita real GDP and standards of living. But growth that falls persistently short validates Thomas Malthus' prediction in his *First Essay on Population* (1798) that population growth would exceed output growth causing starvation and the end of population growth. The big question is: What determines growth in real GDP and real GDP per capita? A theory of economic growth is needed to answer that question.

Observations on the recent history of economic growth raise four questions.

1. What is long-term growth?
2. What are the causes or sources of growth?
3. What are the effects of economic growth, both positive and negative?
4. Can economic policies affect growth?

Although the focus here is mainly on industrial countries, the growth or lack of it in poor countries is also an extremely important issue for the economics of development. To see many other interesting dimensions of growth compared across a much larger sample of countries over longer time periods, visit *Gapminder World* on the web site `www.gapminder.org`.

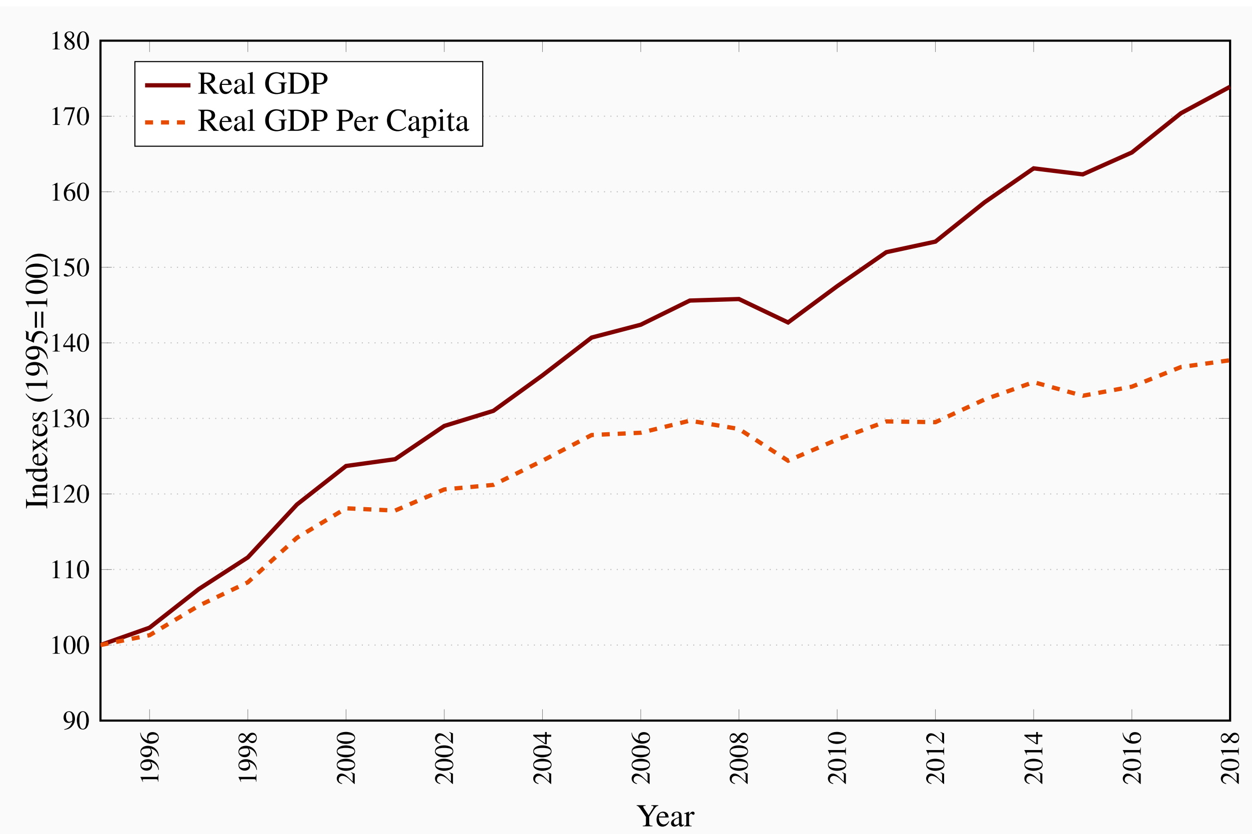
25.1 Patterns of economic growth across countries

Recall, from Chapter 16, that the growth rate of a variable is its percentage change over time. To define economic growth we must specify both the variable to measure and the period over which to measure it. Percentage changes in real GDP and real GDP per person over periods of one year are used to measure annual rates of **economic growth**.

> **Economic growth**: the annual percentage change in real GDP or per capita real GDP.

Annual real GDP measures the total output of final goods and services in the economy for one-year periods. As a result, the annual rate of growth in real GDP is the change in the size of the total economy. But, as discussed in Chapter 16, increases in the size of the total economy may not reflect changes in standards of living or economic welfare. To get an indication of how these grow, stagnate or decline we need to look at growth in real GDP per person. Growth in real GDP can raise standards of living only if it exceeds the growth in population, providing, on average, more goods and services to the average individual.

Figure 25.1: Indexes of real GDP & real GDP per capita: Canada 1995–2018



Source: Statistics Canada Series 17-10-0009-01 and 36-10-0123-01 and author's calculations

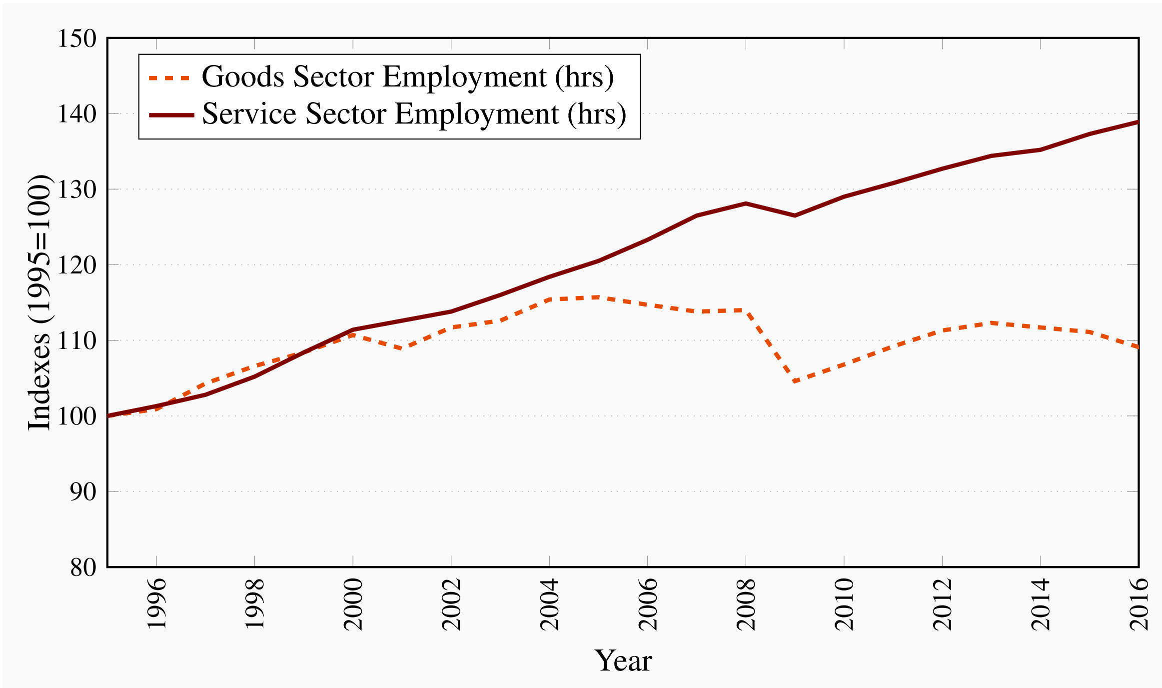
Figure 25.1 illustrates the growth of real GDP and real GDP per capita in Canada from 1995 to 2018. In that time frame the index of real GDP, with 1995 as the base year, increased by about 45 percent from 1995 until the financial crisis of 2008, and then declined during the ‘Great Recession’ before resuming growth in 2010. Over the full 1995–2018 period real GDP grew by about 75 percent.

Population in Canada was also growing in the 1995–2018 period. As a result the growth in per capita real GDP was less than in real GDP. Figure 25.1 shows that growth in real GDP per capita was about 37 percent over the 1995–2018 period. This is a rough measure of the improvement in standard of living for the average Canadian over that time. But there are questions about the distribution of the changes in standard of living.

Structural change in economic output and employment are an integral part of economic growth. In industrial economies evolving supply and demand conditions lead to higher rates of growth in service sector output and employment relative to the goods sector. Different underlying trends in investment in physical and human capital, technology and entrepreneurial experience by sector, together with differences in sector income and price elasticities of demand, explain the longer term evolution of output structures. Table 16.2 showed that in Canada by 2018 the service sector produced 70 percent of real GDP and the goods sector approximately 30 percent.

Structural changes leading to this distribution of output have also had important effects on employment by sector in Canada. Different sector rates of investment and change in technology have resulted in higher productivity growth but slower rates of employment growth in the goods producing sector. Figure 25.2 uses indexes of employment to show the Canadian experience. From 1995 to 2004 goods sector employment increased by about 15 percent and service sector employment by about 18 percent. After that, from 2004 to 2018, goods sector employment fell by about 6 percent. However, service sector employment increased by 18 percent. As a result employment in the goods producing sector fell from about 29 per cent of total employment in 1995 to about 24 per cent in 2018. These changes have implications for the future growth in output, employment, real incomes and the ways in which the per capital real income gains from growth are shared.

Figure 25.2: Indexes of Employment by Industry Sector: Canada 1995–2018



Source: Statistics Canada Table 14-10-0289-01 and author's calculations

Differences in growth rates

Another interesting aspect of growth in real GDP or any other variable is the importance of small differences in compound annual growth rates on growth paths over time. Table 25.1 provides an illustration for selected time periods from one year to 40 years. Annual growth rates in the range of about 0.5% to 4.0% in real GDP and per capita real GDP are common among industrialized countries.

Table 25.1: Importance of small differences in annual growth rates

Growth Rate	**0.5%**	**1.0%**	**2.0%**	**3.0%**	**3.5%**	**4.0%**
Initial value	100	100	100	100	100	100
Year 1	100.5	101.0	102.0	103.0	103.5	104
Year 5	102.5	105.1	110.4	115.9	118.8	121.7
Year 10	105.1	110.5	121.9	134.4	141.5	148.0
Year 20	110.5	122.0	148.6	180.6	199.0	219.1
Year 40	122.1	148.9	220.8	326.2	395.9	480.1

Over a twenty year period an annual growth rate of 2.0 percent would increase per capita real GDP by almost 50 percent, a growth rate of 3.0 percent would increase it by 80 percent and a growth rate of 3.5 percent would double it. Even over a shorter period like the Great Recession and recovery of 2007–2012, a drop in the average annual growth rate from 3.0 percent to 1.0 percent would mean a GDP 10 percent lower than it would have been under steady 3.0 percent growth.

Table 25.2 illustrates the actual variations in growth rates of per capita real GDP across selected industrial economies from 1995 to 2014. Differences in growth rates in 1995–2001 were relatively small and in the range of 2.3 percent and 3.5 percent for most countries in this sample, except for Japan and Korea with rates of 0.5 percent and 4.6 percent respectively. The standard of living was catching up with those in other countries while that in Japan was roughly constant. Except for Japan, all these growth rates by country declined in the next two time periods reported, with particularly large declines as a result of the international financial crisis and recession after 2008. The underlying causes of this slowdown in growth will be examined in terms of a basic theory of economic growth.

Table 25.2: Growth rates in per capita real GDP: Selected countries

	(Average annual % change)		
	1995–2001	**2002–2008**	**2009–2014**
Australia	2.7	1.9	0.9
Canada	2.7	1.6	0.5
France	2.3	1.1	-0.1
Greece	3.3	3.8	-4.5
Japan	0.5	1.5	0.5
Korea	4.6	4.5	2.7
Spain	3.5	1.7	-1.2
Sweden	3.1	2.8	0.3
United Kingdom	2.7	2.4	-0.1
United States	2.6	1.7	0.5

Source: International Monetary Fund, *World Economic Outlook Database*, October 2015, and authors' calculations.

25.2 Growth in potential output

Basic economic growth theory steps back from short-run variations in growth rates. It applies to the **very long run**, a time frame in which wages *and* prices are fully flexible, and the labour force, the stock of capital equipment, knowledge and the technology used in production can change. In this time frame, output fluctuations around potential output are swamped by the growth of potential output itself.

> **Very long run**: the time required for changes to occur in the stock of capital, the size of the labour force, and the technology of production.

An **aggregate production function** describes the links between inputs to production and real GDP produced. Recall that, for the whole economy, Y is real GDP produced by using inputs of labour (N) and capital (K). The function F tells us how much real GDP we get out of particular amounts of labour and capital used in the production process.

$$Y = A \times F(N, K) \tag{25.1}$$

The function $F(...)$ does not change, but changes in N and K cause changes in output Y. Technical progress or improvements in technology are captured separately through A, which measures the state of technology at any date. As technology improves, A increases and more real GDP is produced from the same inputs of labour and capital. A 10 percent increase in A gives 10 percent more real GDP from the same inputs of labour and capital. We describe this as an increase in productivity because outputs per worker and per unit of capital increase. A is often called a measure of total factor productivity (TFP).

> **Aggregate production function**: the link between labour and capital inputs to production, technology and the output as measured by real GDP.

> **Total factor productivity (TFP)**: output relative to the combined inputs of labour and capital, the total factor inputs to production.

Actual real GDP is the output produced at any time based on the actual inputs of capital and labour. In terms of the production function:

- Y_t is real GDP in year t;
- A_t is determined by the current state of technology; and
- K_t and N_t measure the actual use of capital and labour.

These inputs are combined to give output in year t:

$$Y_t = A_t \times F(N_t, K_t) \tag{25.2}$$

Potential output is the real GDP produced when labour and capital are employed at equilibrium rates using the best available technology. A specific production function to recognize this is:

$$Y_P = A_t \times F(N_F, K_0) \tag{25.3}$$

Y_P is potential output produced by operating plants and machinery at their designed capacity (K_0) and using the full employment equilibrium supply of labour services (N_F). A_t is the state of knowledge and technology used in the production process and reflected in the productivity of labour and capital.

Any *growth* in the potential output of goods and services then comes from *growth* in labour inputs to production, *growth* in capital inputs to production, and *changes* in factor productivity as a result of new and improved technology.

Growth accounting measures the sources of growth in real GDP. From the production function, it follows that:

		Effect of		Effect of		Effect of
Growth in	=	Growth in	+	Growth	+	Growth in
Real GDP		Total Factor		in Labour		Capital
		Productivity		Inputs		Inputs

Growth accounting: measurement of the contributions of labour, capital, and technology to growth in output.

The way that growth in capital and labour affects the growth in total output can be measured by the incomes they receive. For example, the income approach to the measurement of net domestic product and GDP identifies these factor income shares in Canada in 2013 as the sum of:

Employee compensation	\$957 billion
Net Corporate surplus	\$240 billion
Net mixed income	\$168 billion
Total factor income	\$1,365 billion

From these data, employment income was about 70 percent of factor income. This is higher than the longer-term average share of employment income in total factor income, but it shows where the measure comes from. Labour's average contribution to and share of national income, measured over time periods of many years, is approximately two-thirds of total factor income. Capital's contribution and share is the remaining one-third of factor income.

The growth in potential GDP over time can then be expressed as the growth in total factor productivity plus the *weighted sum* of the growth in the capital and labour inputs to production as follows:

$$\text{Growth in } Y_P = \text{growth in } A + (2/3) \times (\text{growth in } N) + (1/3) \times (\text{growth in capital stock})$$

This is the basic *growth accounting* equation.

The weights (2/3) and (1/3) applied to growth in labour and capital inputs are based on their shares in national income. They determine the rate of growth in real GDP as a result of growth in the inputs of capital and labour. By these weights, a 10 percent increase in labour input, capital and technology held constant, would result in an increase in real GDP of $2/3 \times 10$ percent, which is 6.6 percent. Similarly, a 10 percent increase in capital input would result in a $1/3 \times 10 = 3.3$ percent increase in real GDP. However, if both N and K grow by 10 percent real GDP grows by 10 percent.

The increase in productivity from improvements in technology cannot be seen and measured directly. As a result, growth accounting classifies these effects as a *residual*. The difference between the growth in real GDP and the weighted sum of the growth in labour and capital inputs is called the **Solow residual**, named after Professor Robert Solow, whose work on growth theory was recognized with a Nobel Prize. The Solow Residual is a measure of the contribution to growth made by improvements in the technology of production that raise the productivity of both labour and capital.

> **Solow residual**: the growth in real GDP or per capita real GDP not caused by growth in factor inputs, but attributed to improved technology.

The Solow residual measured by growth in A is found by rearranging the growth accounting equation as follows:

$$\text{Growth in } A = \text{growth in } Y - (2/3) \times \text{growth in } N - (1/3) \times \text{growth in } K$$

The numerical example in Table 25.3 illustrates the procedure. It assumes data are available for a specific sample period; say 10 years, for the growth rates of real GDP, capital stock, and employment, measured as average annual percentage changes.

Table 25.3: An estimate of the Solow residual using growth accounting

Observed average annual growth rates, %Δ, in:					
Real GDP(Y)	5.0	Employment(N)	2.4	Capital stock(K)	3.9
By growth accounting:					
Growth in $A = 5.0 - (2/3)(2.4) - (1/3)(3.9)$					
Growth in $A = 5.0 - 1.6 - 1.3$					
Contribution of growth in A to growth in real GDP= 2.1 percent					

The calculation made using growth accounting shows that the increased productivity of both labour and capital resulting from improvements in technology was the source of 2.1 percent of the 5.0 percent growth in real GDP in this example.

Earlier research at the Bank of Canada[1] estimated the sources of growth in real GDP and potential GDP in Canada over the period 1950 to 1996 and projections for future growth in potential GDP. The contributions of capital and labour inputs are weighted as in the simple example above. Growth in actual and potential real GDP are the result of growth in factor inputs and the growth in productivity coming from improvements in the technology of production.

That work shows growth in real GDP and potential GDP declined over the 1950 to 1996 period. This slowdown was partly a result of a slowdown in the growth of population, and labour and capital inputs to production. But starting in the 1970s there was also a slowdown in productivity growth, which reduced the rate of growth of output per worker. This slowdown in productivity is examined in more detail later in this chapter. It had important implications for the standard of living in many countries, including Canada. Some further research on productivity growth in Canada by the Centre for the Study of Standards of Living, `www.csls.ca`, and at the Bank of Canada has uncovered another productivity slowdown in Canada relative to the United States after the year 2000, which has continued and is a current source of concern.

25.3 Growth in per capita GDP

Growth in potential GDP measures the increase in the size of the economy, but it does not tell us what is happening to per capita GDP and *standards of living*. To discover the sources of growth in per capita GDP and improvements in standards of living, we need to study the production function in more detail. Then we can use growth accounting to uncover the sources of past growth in per capita GDP.

Consider the same production function we have used for total GDP:

$$Y = A \times F(N,K) (25.1)$$

Again, assume that if technology (A) is constant, real GDP (Y) will grow by 2/3 of the growth in labour input and 1/3 the growth in capital input. If, for example, labour force growth increases employment by 10 percent, with fixed capital stock (K), and technology (A), GDP will increase by 6.67 percent. A similar calculation shows the effects of growth in the capital stock. The weights 1/3 and 2/3 are the *elasticities* of output with respect to the inputs of labour and capital based on factor income shares in national accounts.

Factor contributions and scale economies

The increase in total output when an additional unit of a factor (labour or capital) is used in the production process, and other inputs are held constant, is the **marginal product** of that added factor. The production functions widely used in economics have *diminishing marginal productivity*. As more and more workers are employed using a fixed number of machines, each additional worker adds less and less to total output. The marginal product of labour, the change in total output as a result of using one more worker ($\Delta Y/\Delta N$), falls. Furthermore, because each additional input of labour adds less to total output than the unit before it, output per worker, Y/N, also falls.

[1] J. Kuszczak and R. Dion, “Potential Output Growth: Some Long-Term Projections.” *Bank of Canada Review*, Winter 1997-1998.

Marginal product: the change in total output caused by a change of one unit in the input of that factor to production.

It is often assumed that production involves *constant returns to scale*. Instead of increasing just one input to production, suppose all inputs are increased together, in the same proportions. Labour and capital inputs might both be doubled, for example. Then, if output increases in exactly the same proportions as inputs have increased, there are constant returns to scale.

The production functions used in growth accounting have these properties. Consider the following example. Holding technology constant at $A = 1$ to simplify matters, we can write:

$$Y = N^{2/3} \times K^{1/3} \tag{25.4}$$

using the weights we have used in growth accounting to measure the contributions of labour and capital to output, based on their shares in national income, as the exponents on labour and capital inputs. Table 25.4 gives numerical examples of the way this production function works.

Table 25.4: Changes in outputs as factor inputs change

Production function:			$Y = N^{2/3} \times K^{1/3}$		
Labour input (N)	**Capital input (K)**	**Output (Y)**	$\%\Delta N$	$\%\Delta K$	$\%\Delta Y$
50	20	36.8	–	–	–
55	20	39.3	10.0	0	6.6
55	22	40.5	0	10.0	3.3
60.5	24.2	44.6	10.0	10.0	10.0

The first row of the table shows that a labour input of 50 units combined with a capital input of 20 units gives output:

$$Y = 50^{2/3} \times 20^{1/3} = 36.8 \text{ units}$$

The next three rows illustrate the underlying diminishing returns and constant returns to scale in this production process. An increase of either labour input of 10 percent or capital input of 10 percent with the other input constant increases output, but by less than 10 percent in each case. Because output grows by less than the growth of the input in each case, output per worker or per unit of capital falls.

However, when only one input grows, output per unit of the factor held constant rises. In the second row of the table, capital input is constant when labour input grows. More labour inputs increase total output and output per unit of capital. Similarly, in the third row an increase in capital input increases labour productivity.

The fourth row of the table shows **constant returns to scale**. When labour and capital inputs both increase by the same proportion, output also increases by that proportion, 10 percent in this example. As a result, output per worker and output per unit of capital are constant. In terms of economic growth, equal growth rates of labour input and capital stock make total GDP grow at that same rate, *but leave per capita GDP unchanged*.

> **Constant returns to scale**: equal percentage increases in inputs of labour and capital increase output by the same percentage.

To see the sources of growth in *per capita GDP*, we can manipulate the production function and apply growth accounting. To get per capita GDP, simply divide both sides of the production function in Equation 25.4 by N to give output per worker as follows:

$$Y = A \times N^{2/3} \times K^{1/3} (25.4)$$

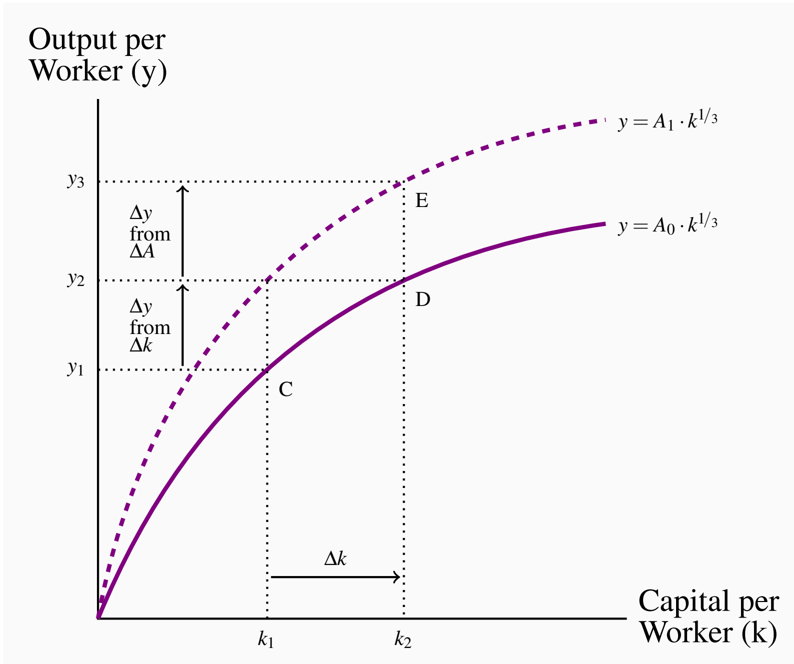
Output per worker $(Y/N) = [A \times (N^{2/3} \times K^{1/3})$ divided by $(N)]$, the number of workers employed.

To make the notation a bit neater, we can use lower case letters to indicate *output per worker* $(y = Y/N)$ and *capital per worker* $(k = K/N)$. This gives:

$$y = A \times k^{1/3} \tag{25.5}$$

Figure 25.3 illustrates this production function in a diagram. The ratio of capital stock to labour, k, is measured on the horizontal axis. Output per worker, y, is measured on the vertical axis. Two per-worker production functions are used to distinguish between the effects of increases in capital stock and the effects of improvements in technology.

Figure 25.3: The effects of increases in the capital labour ratio and improvements in technology on output per worker



An increase in the capital to labour ratio from k_1 to k_2 raises output per worker from y_1 to y_2. The shape of the production function shows that further increases in k will give further but smaller increases in y. A change in technology, ΔA, shifts the production function up as y increases at every k. The combined effects of an increase in k and an increase in A are increased output per worker from y_1 to y_3.

The declining slopes of both production functions illustrate the diminishing returns that lead to the smaller and smaller changes in output per worker as the capital/labour ratio increases. For example, starting at point C, an increase in the ratio of capital to labour moves the economy along the production function to point D. Output per worker increases at a decreasing rate. This shows that increased capital to labour ratios can increase output per worker until diminishing returns set in and limit sustained increases in output per worker. **Sustained growth in per capita real GDP**, in this basic model of growth, depends on improvements in technology to overcome the diminishing returns to increases in the capital to labour ratio.

> **Sustained growth in per capita real GDP**: improvements in technology overcome the diminishing returns to increases in the capital to labour ratio.

An improvement in technology that increases productivity ($A_2 > A_1$) shifts the production function up. At capital to labour ratio k_2, for example, the increased productivity moves the economy from D to E as output per worker rises from y_2 to y_3. This shows that growth in per capita output, moving from y_1 to y_3, points C to E in Figure 25.3, is a result of both the growth in the capital to labour ratio and improvements in productivity.

Once again, growth accounting allows us to sort out the effects of these two factors on growth in output per worker and per capita GDP. Table 25.5 uses Canadian experience from 1990 to 2013 as an example.

Table 25.5: Sources of growth in per-worker GDP in Canada, 1990–2013

	Real GDP Y **($B)**	**Employ** N **(M)**	**Capital Stock** K **($B)**	**Real GDP per worker** $Y/N = y$ **($K)**	**Capital per worker** K/N **($K)**	**Growth in** Y/N $\%\Delta y$	**Growth in** K/N $\%\Delta k$	**Contribution from** $\Delta k =$ $\%\Delta k/3$	**Solow Residual** $\%\Delta A$
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)=(7)/3**	**(9)=(6)−(8)**
1990	993	13.1	1,427	75.98	109.18	–	–	–	–
1995	1,082	13.3	1,491	81.54	112.36	7.3	2.9	1.0	6.4
2000	1,328	14.8	1,603	89.97	108.60	10.3	−3.3	−1.1	11.5
2005	1,502	16.2	1,739	92.83	107.48	3.2	−1.0	−0.3	3.5
2013	1,698	17.7	2,094	95.99	118.37	3.4	10.1	3.4	0.0

Source: Statistics Canada Tables: 36-10-104-01, 14-10-0287-03, 36-10-0098-01.

The first three columns of the table give data on real GDP, employment, and real capital stock for four selected years over the 1990 to 2013 period. Columns **(4)** and **(5)** use these data to calculate the output per worker and capital per worker that we see in our per-worker production function. The growth in output per worker and the growth in capital per worker are reported in columns **(6)** and **(7)** as the percentage changes over three five-year periods and a final eight year period to use the latest data available.

Growth accounting divides the sources of growth in output per worker between increases in capital per worker and increases in productivity based on improvements in technology. From the production function, we know that an increase in the capital/labour ratio increases output per worker by a factor of 1/3. Column **(8)** in the table reports this weighted contribution of the increase in capital per worker to the increase we see in output per worker. Subtracting these contributions from the increases in output per worker gives the Solow residual, column **(9)**, which again is a measure of the effect of improvements in technology on output per worker.

The 1990 to 2013 period is of interest because the Canadian experience provides different examples of growth in GDP per worker. In the first five years, 1990 to 1995, there was very little growth in employment but some growth in capital stock. As a result, the capital to labour ratio, k, increased by 2.9 percent and accounted for 1.0 percentage points, or 13 percent of the 7.3 percent growth in real GDP per worker. Improved technology as measured by the Solow residual contributed the other 6.4 percentage points.

By contrast, employment and capital stock both grew strongly from 1995 to 2000, but employment growth (11.2 percent) exceeded capital stock growth (7.6 percent). As a result, the growth in the capital to labour ratio was negative, at −3.3 percent. Nevertheless, output per worker grew more in the second period, up 10.3 percent. Productivity gains from improved technology, again measured by the Solow residual, were the major source of strong growth in GDP per worker.

The decline in the capital to labour ratio in the 1995 to 2000 period appears to have reduced the growth in output per worker in the 2000 to 2005 period. Even though capital stock increased by about 8 percent and employment grew by 9.5 percent, output per worker increased by only 3.2 percent over the period. The Solow residual in column **(9)** shows a very small contribution to growth in output per worker from improved technology.

In the final period, 2005–2013 the calculated Solow residual is zero. A strong increase in capital stock and the capital labour ratio explains all the observed growth in output per worker.

Thinking of these experiences in terms of the production functions in Figure 25.3 illustrates the differences between sub-periods. From 1990 to 1995, the economy moved to the right along the production function as k grew by 2.9 per cent, for example, from C to D in the diagram. This provided a 1.0 percent increase in y, as from y_1 to y_2. Improved technology shifted the production function up to further increase output per worker from y_2 to y_3 at point E in the diagram. In the next period, 1995 to 2000, the movement along the production function was in the opposite direction, to the left from E, as k declined. However, a very strong effect from improved technology as measured by the Solow residual of 11.5 percent shifted the production function upward (not shown in the diagram) and sustained the growth in output per worker at 10.3 percent.

These examples and the discussion of the sources of growth emphasize two key aspects of the growth process. One aspect is the growth in the stock of capital, which comes from the flow of savings and investment in the economy. The other is changing technical knowledge and technology of production. These are the keys to sustained growth in total output and standards of living, but their sources are more obscure than the sources of growth in capital stock. Indeed, pessimism about the fate of society was based on both the inadequacy of investment and stagnant technology.

25.4 Technology & growth in real per capita GDP

Advances in knowledge based on research and development and experience are the key to sustained rates of productivity growth and improvements in standards of living.

Technical knowledge

Every society has some level of technical knowledge about production practices. Part of this knowledge is recorded in technical documents, books, and plans. But it develops and is often captured in current working practices based on experience. This technical knowledge grows through **invention** that uncovers new understanding and knowledge. What usually follows is **innovation** that applies new knowledge to actual production techniques.

> **Invention**: the discovery of new knowledge.
>
> **Innovation**: the application of new knowledge into production techniques.

Industrial and economic history is often written in terms of the sequence of major inventions and innovations. Improvements in transportation from the wheel, to steam engines, internal combustion engines and aircraft transformed the size of the market and the degree of specialization and trade. The generation and transmission of electricity provided a new, more efficient and flexible source of light and power leading to improved communications and data management.

Historically, the agriculture 'revolution' was based on the application of science and technology to products and production processes. That evolution continues: New organization and techniques increased productivity; new machinery further increased labour productivity and yielded economies of scale; and new science provided better seed, fertilizer and fuels to power new mechanical equipment. Increases in agriculture productivity based on these and other advances reduced the labour and land required to supply food to the population. At the same time, increased industrialization increased the demand for labour. Economies became more sector specialized with identifiable primary, manufacturing and service sectors that worked together to improve productivity and standards of living.

Additions to human capital were as important as increases in physical capital in this growth process. Human capital is knowledge and experience with production processes. Experience improves workers' efficiency. Higher levels of education and training lets workers use more complex production equipment and techniques and contribute to cost improvements in the organization of production. In short, productivity is enhanced by the complementarity and synergy between human and physical capital.

The role of research and development

The invention and innovation that lies behind productivity improvements has many sources. Familiarity with a product, process or production technique often suggests a better way to work, or a better product design. Experience and frustration can lead to curiosity and invention. However, most invention and innovation has its origin in specialized research and development. From this perspective the output of new ideas depends on the resources allocated to R&D. There are costs involved and the focus and usefulness of the results may not be apparent for some time, if ever. Some research is pure research that seeks new knowledge and understanding. Most of it takes place in university departments, usually funded in part by industry, in part by government and in part by private individuals. Applied research, on the other hand is usually based in and financed by industrial firms, with some support from government grants and tax incentives. New knowledge with commercial applications is the goal of this applied R&D.

Investment in R&D, like investment in human and fixed capital, is risky. It cannot be known in advance that new, useful knowledge will result. Funding is provided and committed in the hope and expectation that research projects will be successful. Even if research succeeds it may be difficult to apply the new knowledge or technology in a way that increases your market power enough to allow recovery of the costs. Success gives competitors strong incentives to copy the new product or create a close substitute for it. Patent laws and government subsidies are designed to help private companies and individuals recover their costs of R&D, recognizing that society benefits from improvements in technology.

25.5 Recent growth studies and policy issues

Basic growth theory and the basic growth accounting methodology leave a large part of aggregate growth and productivity growth in the Solow Residual and unexplained. Growth in employment and capital stock, and the relationship between them, play important roles in aggregate growth and labour productivity growth, but the effects of changes in the characteristics of labour and the composition of capital stock reside in the total factor productivity estimates of the Solow residual. This residual captures changes in technology along with other undefined factors.

Recent work on productivity growth still uses the growth accounting methodology, but focused on particular sectors of the economy and with extensions designed to unpack some things previously left in the residual. These include, in particular, investment in different components of the capital stock and changes in the composition and quality of labour.

Increases in capital per worker or per hour of work—capital deepening—continues to be important, but increases in capital stock are disaggregated into investment into several categories such as:

- Information and communications technology;
- Machinery and equipment; and
- Physical structures.

Changes in educational qualifications, gender structure, and age structure are used as indicators of changes in the composition and quality of the labour force. To the extent that these measures of change in characteristics and structure affect productivity less remains in the residual. More importantly, these are areas in which policies to support education, training, and labour force participation could affect productivity growth. Table 25.6 gives an example of some recent results based on this approach.

The estimates in the table show the effect of increases in the information, communications, and technology components of the capital stock. Particularly over the 1997 to 2000 period, the increase in ICT capital per unit of labour input was the largest part of capital deepening. This is also the period of strongest growth in labour productivity and the largest increase in total factor productivity. The latter captures the effects of other improvements in technology.

Table 25.6: Sources of Canadian business sector labour productivity growth, 1974–2005 (%/yr)

	1974–1996	1997–2000	2000–2005
Labour productivity	1.4	3.0	1.0
Capital deepening	1.1	1.0	0.7
Info & communications technology (ICT)	0.4	0.7	0.3
Non-ICT	0.7	0.4	0.4
Labour quality	0.4	0.4	0.4
Total factor productivity	0.0	1.6	–0.1

Source: R. Dion, "Interpreting Canada's Productivity Performance in the Past Decade: Lessons from Recent Research." Bank of Canada Review, Summer 2007.

The estimated contributions of changes in labour quality to labour productivity growth provide an interesting refinement of the growth accounting methodology. Ignoring this change, as the simple accounting process did in earlier examples, leaves the contribution of changes in the structure and characteristics of the labour force in the residual measure of total factor productivity. The results in the table extract and quantify this important source of productivity growth.

Changes in the structure of investment and in the quality of labour are both areas in which public policy plays a role. Government tax policy can be designed to encourage producers to direct investment to areas like communications and information technology. Some of this investment has effects confined to specific industries. Other parts create and strengthen national capacity and efficiency that is widely available and used by business and households, for example high-speed internet service and expanded wireless service and capacity. Educational policy and other human resource policies contribute to the quality, adaptability, and mobility of the labour force.

Endogenous growth theory

Another interesting strand of work makes more fundamental changes in growth theory. The simple neoclassical growth theory made economic growth depend on **exogenous variables**, the rate of population growth, the saving rate, and the rate of capital accumulation, whose values are determined outside the growth model. The subsequent work on catch-up and convergence makes technical progress respond to economic and political factors. But it would be nice to have a stronger link between economic behaviour and the rate of economic growth. We want to make *growth endogenous*, or determined within our model. **Endogenous growth** implies that the steady-state growth rate is affected by economic behaviour and economic policy.

> **Exogenous variable**: a variable with a value determined outside the model.

> **Endogenous growth**: growth determined economic behaviour and policy within the model.

Professor Paul Romer of New York University received the Nobel Prize in Economics for his pioneering work on endogenous growth theory. Growth theories are built on the saving that drives investment and capital accumulation. While the production function used the basic growth model, increasing the stock of capital and the capital labour ratio leads to a diminishing marginal product of capital. Output and output per worker increase at a decreasing rate but in the steady state output grows at the rate of growth of population and labour force. Output per worker cannot grow indefinitely.

Endogenous growth theory is based on a different view of the roles capital plays in the growth process. It retains the assumption that, for the individual firm, the marginal product of capital decreases as capital stock increases. But it adds the insight that in the *aggregate economy* there may be significant positive externalities to the increase in capital stock. In other words increases in capital stock by one firm may result in improved productivity and output in other firms. Improved internet speed and access, or faster wireless voice and data transmission comes from investment by internet service providers and wireless companies. Using the new capacity provided improves the efficiency and productivity of production in other companies. Aggregate productivity and output grows as a result of increased investment by individual producers.

This line of argument is illustrated by the shapes of the production functions. For the individual firm the production function still looks like that in Figure 25.1, concave to the horizontal axis. Increases in the capital/labour ratio increase output per worker at a decreasing rate. For the *aggregate economy*, according to endogenous growth theory, increases in capital per worker do not face diminishing returns because of the externalities that come with increases in the capital stock. As a result, the *aggregate production function has constant returns* to capital. It is a straight line rising from the origin with a constant slope. Growth in the capital/labour ratio produces a constant rate of growth in output per worker and aggregate output. A 10 percent increase in capital per worker gives a 10 percent increase in output per worker.

Constant returns to increases in the *aggregate* capital to labour ratio in the economy allows an escape from the key growth limitation in the neoclassical theory. It makes growth endogenous and dependent upon parameters that could be influenced by private behaviour or public policy. Any policy that succeeded in raising the rate of investment would permanently raise the growth rate. Similarly, any policy achieving a one-time improvement in technology (for example, greater workplace efficiency) would permanently raise the growth rate of capital per worker. This would mean permanently faster output growth.

Not only can government policy affect growth in this framework, government intervention may also increase efficiency. In the simple endogenous model described here, there are externalities to capital accumulation. Individual producers may not realize that by investing to increase their capital stock they may also improve productivity in other firms. Public policy that recognizes this economy wide effect can subsidize investment to increase investment and aggregate economic growth. By the same argument, externalities to investment in human capital support government subsidies to education and training.

However, endogenous growth theory faces criticism based on the assumption that there are exactly constant returns in the aggregate from accumulating one factor of production. The diminishing returns in the basic model make long-run growth exogenous. An economy with *increasing* returns experiences continuously increasing capital stock and output. This sort of explosive growth does not correspond to any empirical observation.

New growth theory emphasizes the roles of research and development, innovation, education and "learning by doing" as sources of improved technology and productivity. Research, development, and innovation come from decisions to invest in new knowledge and to apply it to production processes. Education comes from decisions to invest in human capital. Learning by doing is a natural outcome of employment experience. All these are ongoing processes, although they may be pursued unevenly over time as economic conditions and economic policies change.

As a result, increases in the stock of capital and the level of employment always embody new technology and knowledge. There is no separation between increases in the capital to labour ratio and the state of technology, as in the basic neoclassical model. But if new capital stock and new employees bring new technology to the production process, the per worker production function *shifts up* as capital per worker increases. As capital per worker increases, the economy moves up and along a new per worker production function because the embodied technological improvements offset otherwise diminishing returns.

Recent studies of the sources of productivity growth based on growth accounting are consistent with this approach to reconciling neoclassical and new growth theory. The findings reported in Table 25.6 above show the contributions to productivity growth made by different types of capital equipment and changes in labour force structure. These changes in "technology" are made integral parts of the growth in capital stock and employment rather than left as exogenous residuals.

The costs of growth

Can the benefits of economic growth be outweighed by its costs? Pollution, congestion, and a hectic lifestyle are a high price to pay for more cars and trucks, washing machines, video games, smart phones and tablets.

Since GDP is an imperfect measure of the true economic value of goods and services produced by the economy, there is no presumption we should want to maximize the growth of measured GDP. Without government intervention, a free market economy produces too much pollution. But the elimination of all pollution is also wasteful. Society should undertake activities accompanied by pollution up to the point at which the marginal net benefit of the goods produced equals the marginal pollution cost imposed on society. Government intervention, through pollution taxes or regulation by environmental standards, can move the economy towards a more efficient allocation of resources and a higher standard of living, broadly defined.

The full implementation of such a policy would (optimally) reduce growth of measured GDP to below the rate when there is no restriction on pollution and congestion. This is the most sensible way in which to approach the problem. It tackles the issues directly.

In contrast, the "zero growth" solution is a blunt instrument. It does not differentiate measured outputs that have social costs, from outputs without new social costs. As a result there are no new incentives to minimize the externalities already caused by pollution, congestion and environmental degradation. These call for taxes or incentives to reduce current social costs. 'No growth' does not go far enough.

A more extensive and inclusive measure of GDP would be a step toward recognizing and dealing with the costs of growth. Such a measure might include both positive aspects of production and consumption that contribute to welfare, such as environmental quality and low levels of congestion. It might also adjust to give a more complete measure of both private and social costs of different patterns of production, consumption and leisure. We have seen one way to approach this in the United Nations Human Development Index in Chapter 16, but a broader based index is needed.

Conclusion

This completes the introduction to the theory and modelling needed to study the major performance and policy issues in the macro economy. This approach has integrated household and business expenditure decisions, government taxation, expenditure and budget balances, and the monetary, financial sectors and monetary policy, to explain the causes and effects of business cycle fluctuations in national output and employment. It also defines and explains monetary and fiscal policy stabilization roles for the central bank and the government. The brief discussion of the theory of economic growth puts a longer-term perspective on the changes in standards of living over time. Perhaps most importantly, modelling the macro economy emphasizes the complex linkages and interdependencies that determine the modern industrial economy's responses to aggregate demand and supply disturbances.

Key Concepts

Economic growth is the percentage annual increase in real GDP or per capita real GDP. It is an imperfect measure of the rate of increase of economic well-being because of the limitations of the measurement of GDP.

Very long run: the time required for changes to occur in the stock of capital, the size of the labour force, and the technology of production.

Total factor productivity (TFP): output relative to the combined inputs of labour and capital, the total factor inputs to production.

Growth accounting: a method of measuring the contributions of growth in inputs of labour and capital and the state of technology to overall growth.

Solow residual: the growth in real GDP or per capita real GDP not caused by growth in factor inputs, but attributed to improved technology.

Growth in per capita real GDP has **two main sources**, namely, growth in the ratio of capital to labour in the production process, and improvements in technology. Growth accounting provides a way of measuring the sources of growth in per capita real GDP.

Marginal product: the change in total output caused by a change of one unit in the input of that factor to production.

Sustained growth in per capita real GDP: improvements in technology overcome the diminishing returns to increases in the capital to labour ratio.

Recent research on the **growth of Canada's potential GDP** found that the contribution of productivity growth from technology declined in the 1980s and early 1990s and again from 2000 to 2005.

Invention: the discovery of new knowledge.

Innovation: the application of new knowledge into production techniques.

Exogenous variable: a variable with a value determined outside the model.

Endogenous growth: growth determined economic behaviour and policy within the model.

Exercises for Chapter 25

Exercise 25.1

(a) What is the distinction between growth in potential GDP and growth in per capita real GDP?
(b) Why is this distinction important to an evaluation of the relationship between economic growth and growth in standards of living?
(c) Which grows more rapidly, potential GDP or per capita real GDP?

Exercise 25.2 Consider two countries with the same level of potential GDP, say $100 billion, today. Suppose potential GDP grows at an annual rate of 3.5 percent (0.035) in one country and 3.25 percent (0.0325) in the second country. Based on this information:

(a) What do you predict for the percentage difference in potential GDP between the two countries 10 years in the future?
(b) 20 years in the future? [Note that the growth rates will compound to determine real GDP according to the following formula: $Y_t = Y_0(1 + \text{growth rate})^t$.]

Exercise 25.3 Suppose you have the following information about an economy:

Average annual rates of growth from 1998 to 2008:

Potential GDP	3.5%
Labour force	2.1%
Capital stock	3.0%

Share of labour income in national income: 2/3. Using growth accounting, find the contribution to the annual growth in potential GDP that came from:

(a) Growth in labour force
(b) Growth in capital stock
(c) Improved productivity as measured by the Solow residual.

Exercise 25.4 If technology were constant while labour force grew at a rate of 2.5% a year, capital stock grew at 1.5% per year and the share of labour income in national income was 70%, how fast would potential GDP grow?

Exercise 25.5 Suppose you have the following information for two economies:

		Country A	Country B
Average annual	**i. Labour force**	2.5%	4.0%
growth rates:	**ii. Capital stock**	3.5%	3.5%
Labour income/national income:		2/3	2/3

(a) Assuming a constant state of technology, which of these two countries will have the faster rate of growth in total real GDP?

(b) Which of the two countries will have the faster rate of growth in per capita real GDP?

(c) What differences, if any, do you see in the growth rates of the capital to labour ratios in the two countries?

(d) Explain the reasons for the differences in growth rates you have found.

Exercise 25.6 In Wonderland, labour force and capital stock both grow at the rate of 2.5% a year but technology is constant. At what rate will potential GDP grow? At what rate will per capita GDP grow? If improvements in technology increased total factor productivity by 1.5% year, how fast would per capita real GDP grow?

Exercise 25.7

(a) Why do economists emphasize that improvements in technology are the key to improvements in standards of living?

(b) Using a diagram that shows the relationship between capital per worker and output per worker, illustrate and explain why growth in capital per worker cannot provide sustained growth in output per worker and standards of living.

(c) In the diagram in part (b), show how an improvement in productivity coming from improved technology could provide sustained increases in standards of living.

Glossary: Microeconomic Terms

Accounting profit: is the difference between revenues and explicit costs. (7.2)

Adverse selection occurs when incomplete or asymmetric information describes an economic relationship. (14.1)

Affordable set of goods and services for the consumer is bounded by the budget line from above; the **non-affordable set** lies strictly above the budget line. (6.3)

Age-earnings profiles define the pattern of earnings over time for individuals with different characteristics. (13.2)

Agent: usually a manager who works in a corporation and is directed to follow the corporation's interests. (7.2)

Allocative inefficiency arises when resources are not appropriately allocated and result in deadweight losses. (10.4)

Asset price: the financial sum for which the asset can be purchased. (12.4)

Asymmetric information is where at least one party in an economic relationship has less than full information and has a different amount of information from another party. (14.1)

Autarky denotes the no-trade situation. (15.3)

Average fixed cost is the total fixed cost per unit of output. (8.4)

Average product of labour is the number of units of output produced per unit of labour at different levels of employment. (8.3)

Average revenue is the price per unit sold. (10.2)

Average total cost is the sum of all costs per unit of output. (8.4)

Average variable cost is the total variable cost per unit of output. (8.4)

Bid rigging is an illegal practice in which bidders (buyers) conspire to set prices in their own interest. (14.5)

Boom: a period of high growth that raises output above normal capacity output. (1.6)

Break-even price corresponds to the minimum of the ATC curve. (9.3)

Budget constraint defines all bundles of goods that the consumer can afford with a given budget. (6.3)

Capital gains (losses) arise from the ownership of a corporation when an individual sells a share at a price higher (lower) than when the share was purchased. (7.1)

Capital market: a set of financial institutions that funnels financing from investors into bonds and stocks. (7.4)

Capital services are the production inputs generated by capital assets. (12.4)

Capital stock: the buildings, machinery, equipment and software used in producing goods and services. (1.6)

Carbon taxes are a market-based system aimed at reducing GHGs. (5.7)

Cardinal utility is a measurable concept of satisfaction. (6.2)

Cartel: is a group of suppliers that colludes to operate like a monopolist. (10.6)

Cluster: a group of firms producing similar products, or engaged in similar research. (8.8)

Collusion is an explicit or implicit agreement to avoid competition with a view to increasing profit. (11.4)

Comparative static analysis compares an initial equilibrium with a new equilibrium, where the difference is due to a change in one of the other things that lie behind the demand curve or the supply curve. (3.4)

Complementary goods: when a price reduction (rise) for a related product increases (reduces) the demand for a primary product, it is a complement for the primary product. (3.4)

Concentration ratio: N-firm concentration ratio is the sales share of the largest N firms in that sector of the economy. (11.2)

Conjecture: a belief that one firm forms about the strategic reaction of another competing firm. (11.4)

Content requirement: requires that a specified percentage of the final value of a product originate in the producing economy. (15.4)

Constant returns to scale implies that output increases in direct proportion to an equal proportionate increase in all inputs. (8.6)

Consumer equilibrium occurs when marginal utility per dollar spent on the last unit of each good is equal. (6.2)

Consumer optimum occurs where the chosen consumption bundle is a point such that the price ratio equals the marginal rate of substitution. (6.3)

Consumer price index: the average price level for consumer goods and services. (2.1)

Consumer surplus is the excess of consumer willingness to pay over the market price. (5.2)

Consumption possibility frontier (CPF): the combination of goods that can be consumed as a result of a given production choice. (1.4)

Consumption possibility frontier defines what an economy can consume after production specialization and trade. (15.3)

Corporation or company is an organization with a legal identity separate from its owners that produces and trades. (7.1)

Corrective tax seeks to direct the market towards a more efficient output. (5.5)

Cournot behaviour involves each firm reacting optimally in their choice of output to their competitors' decisions. (11.5)

Credible threat: one that, after the fact, is still optimal to implement. (11.6)

Cross-price elasticity of demand is the percentage change in the quantity demanded of a product divided by the percentage change in the price of another. (4.4)

Cross-section data: values for different variables recorded at a point in time. (2.1)

Data: recorded values of variables. (2.1)

Deadweight loss of a tax is the component of consumer and producer surpluses forming a net loss to the whole economy. (5.4)

Decreasing returns to scale implies that an equal proportionate increase in all inputs leads to a less than proportionate increase in output. (8.6)

Demand is the quantity of a good or service that buyers wish to purchase at each possible price, with all other influences on demand remaining unchanged. (3.2)

Demand curve is a graphical expression of the relationship between price and quantity demanded, with other influences remaining unchanged. (3.3)

Demand for labour: a derived demand, reflecting the demand for the output of final goods and services. (12.1)

Demand is elastic if the price elasticity is greater than unity. It is **inelastic** if the value lies between unity and 0. It is **unit elastic** if the value is exactly one. (4.1)

Depreciation is the annual change in the value of a physical asset. (12.4)

Differentiated product is one that differs slightly from other products in the same market. (11.3)

Diminishing marginal rate of substitution reflects a higher marginal value being associated with smaller quantities of any good consumed. (6.3)

Diminishing marginal utility implies that the addition to total utility from each extra unit of a good or service consumed is declining. (6.2)

Discrimination implies an earnings differential that is attributable to a trait other than human capital. (13.6)

Distortion in resource allocation means that production is not at an efficient output. (5.4)

Diversification reduces the total risk of a portfolio by pooling risks across several different assets whose individual returns behave independently. (7.4)

Dividends are payments made from after-tax profits to company shareholders. (7.1)

Dominant strategy: a player's best strategy, whatever the strategies adopted by rivals. (11.4)

Dumping is a predatory practice, based on artificial costs aimed at driving out domestic producers. (15.6)

Duopoly defines a market or sector with just two firms. (11.1)

Dynamic gains: the potential for domestic producers to increase productivity by competing with, and learning from, foreign producers. (15.4)

Economic (supernormal) profits are those profits above normal profits that induce firms to enter an industry. Economic profits are based on the opportunity cost of the resources used in production. (9.4)

Economic efficiency defines a production structure that produces output at least cost. (8.1)

Economic equity is concerned with the distribution of well-being among members of the economy. (2.3)

Economic profit: is the difference between revenue and the sum of explicit and implicit costs. (7.2)

Economies of scope occur if the unit cost of producing particular products is less when combined with the production of other products than when produced alone. (8.8)

Economy-wide PPF is the set of goods combinations that can be produced in the economy when all available productive resources are in use. (1.5)

Education premium: the difference in earnings between the more and less highly educated. (13.2)

Efficiency addresses the question of how well the economy's resources are used and allocated. (5.1)

Efficient market: maximizes the sum of producer and consumer surpluses. (5.3)

Efficient supply of public goods is where the marginal cost equals the sum of individual marginal valuations, and each individual consumes the same quantity. (14.1)

Elasticity of supply is defined as the percentage change in quantity supplied divided by the percentage change in price. (4.6)

Equilibrium price: equilibrates the market. It is the price at which quantity demanded equals the quantity supplied. (3.2)

Equity deals with how society's goods and rewards are, and should be, distributed among its different members, and how the associated costs should be apportioned. (5.1)

Excess burden of a tax is the component of consumer and producer surpluses forming a net loss to the whole economy. (5.4)

Excess demand exists when the quantity demanded exceeds quantity supplied at the going price. (3.2)

Excess supply exists when the quantity supplied exceeds the quantity demanded at the going price. (3.2)

Exclusive sale: where a retailer is obliged (perhaps illegally) to purchase all wholesale products from a single supplier only. (14.5)

Explicit costs: are the measured financial costs. (7.2)

Externality is a benefit or cost falling on people other than those involved in the activity's market. It can create a difference between private costs or values and social costs or values. (5.5)

Firm-specific skills raise a worker's productivity in a particular firm. (13.3)

Fixed costs are costs that are independent of the level of output. (8.4)

Flow is the stream of services an asset provides during a period of time. (12.4)

Full employment output $Y_c = (\text{number of workers at full employment}) \times (\text{output per worker})$. (1.6)

Game: a situation in which contestants plan strategically to maximize their profits, taking account of rivals' behaviour. (11.4)

General skills enhance productivity in many jobs or firms. (13.3)

Gini index: a measure of how far the Lorenz curve lies from the line of equality. Its maximum value is one; its minimum value is zero. (13.7)

Globalization is the tendency for international markets to be ever more integrated. (8.7)

Greenhouse gases that accumulate excessively in the earth's atmosphere prevent heat from escaping and lead to global warming. (5.7)

Gross investment is the production of new capital goods and the improvement of existing capital goods. (12.4)

High (low) frequency data series have short (long) intervals between observations. (2.1)

Human capital is the stock of expertise accumulated by a worker that determines future productivity and earnings. (13.1)

Imperfectly competitive firms face a downward-sloping demand curve, and their output price reflects the quantity sold. (11.1)

Implicit costs: represent the opportunity cost of the resources used in production. (7.2)

Income elasticity of demand is the percentage change in quantity demanded divided by a percentage change in income. (4.5)

Increasing (decreasing) cost industry is one where costs rise (fall) for each firm because of the scale of industry operation. (9.5)

Increasing returns to scale implies that, when all inputs are increased by a given proportion, output increases more than proportionately. (8.6)

Index number: value for a variable, or an average of a set of variables, expressed relative to a given base value. (2.1)

Indifference curve defines combinations of goods and services that yield the same level of satisfaction to the consumer. (6.3)

Indifference map is a set of indifference curves, where curves further from the origin denote a higher level of satisfaction. (6.3)

Industry supply (short run) in perfect competition is the horizontal sum of all firms' supply curves. (9.3)

Industry supply in the long run in perfect competition is horizontal at a price corresponding to the minimum of the representative firm's long-run ATC curve. (9.5)

Inferior good is one whose demand falls in response to higher incomes. (3.4)

Inferior goods have a negative income elasticity. (4.5)

Inflation (deflation) rate: the annual percentage increase (decrease) in the level of consumer prices. (2.1)

Intermediate good: one that is used in the production of final output. (15.4)

Intra-firm trade is two-way trade in international products produced within the same firm. (15.4)

Intra-industry trade is two-way international trade in products produced within the same industry. (15.4)

Invention is the discovery of a new product or process through research. (10.7)

Labour force: is that part of the population either employed or seeking employment. (12.2)

Law of demand states that, other things being equal, more of a good is demanded the lower is its price. (6.2)

Law of diminishing returns: when increments of a variable factor (labour) are added to a fixed amount of another factor (capital), the marginal product of the variable factor must eventually decline. (8.3)

Learning by doing can reduce costs. A longer history of production enables firms to accumulate knowledge and thereby implement more efficient production processes. (8.8)

Limited liability means that the liability of the company is limited to the value of the company's assets. (7.1)

Long-run: a period of time that is sufficient to enable all factors of production to be adjusted. (8.2)

Long-run average total cost is the lower envelope of all the short-run ATC curves. (8.6)

Long-run equilibrium in a competitive industry requires a price equal to the minimum point of a firm's ATC. At this point, only normal profits exist, and there is no incentive for firms to enter or exit. (9.4)

Long-run marginal cost is the increment in cost associated with producing one more unit of output when all inputs are adjusted in a cost minimizing manner. (8.6)

Longitudinal data follow the same units of observation through time. (2.1)

Lorenz curve describes the cumulative percentage of the income distribution going to different quantiles of the population. (13.7)

Luxury good or service is one whose income elasticity equals or exceeds unity. (4.5)

Macroeconomics studies the economy as a system in which feedback among sectors determine national output, employment and prices. (1.1)

Marginal abatement curve reflects the cost to society of reducing the quantity of pollution by one unit. (5.7)

Marginal cost of production is the cost of producing each additional unit of output. (8.4)

Marginal damage curve reflects the cost to society of an additional unit of pollution. (5.7)

Marginal product of capital is the output produced by one additional unit of capital services, with all other inputs being held constant. (12.5)

Marginal product of labour is the addition to output produced by each additional worker. It is also the slope of the total product curve. (8.3)

Marginal rate of substitution is the slope of the indifference curve. It defines the amount of one good the consumer is willing to sacrifice in order to obtain a given increment of the other, while maintaining utility unchanged. (6.3)

Marginal revenue is the additional revenue accruing to the firm resulting from the sale of one more unit of output. (9.3)

Marginal revenue is the change in total revenue due to selling one more unit of the good. (10.2)

Marginal revenue product of labour is the additional revenue generated by hiring one more unit of labour where the marginal revenue declines. (12.1)

Marginal utility is the addition to total utility created when one more unit of a good or service is consumed. (6.2)

Market demand: the horizontal sum of individual demands. (3.8)

Market failure defines outcomes in which the allocation of resources is not efficient. (14.1)

Microeconomics is the study of individual behaviour in the context of scarcity. (1.1)

Minimum efficient scale defines a threshold size of operation such that scale economies are almost exhausted. (8.6)

Mixed economy: goods and services are supplied both by private suppliers and government. (1.1)

Model is a formalization of theory that facilitates scientific inquiry. (1.2)

Monopolist: is the sole supplier of an industry's output, and therefore the industry and the firm are one and the same. (10.1)

Monopolistic competition defines a market with many sellers of products that have similar characteristics. Monopolistically competitive firms can exert only a small influence on the whole market. (11.1)

Monopolistically competitive equilibrium: in the long run requires the firm's demand curve to be tangent to the ATC curve at the output where $MR = MC$. (11.3)

Monopsonist is the sole buyer of a good or service and faces an upward-sloping supply curve. (12.1)

Moral hazard may characterize behaviour where the costs of certain activities are not incurred by those undertaking them. (14.1)

Nash equilibrium: one in which each player chooses the best strategy, given the strategies chosen by the other player, and there is no incentive for any player to move. (11.4)

Natural monopoly: one where the ATC of producing any output declines with the scale of operation. (10.1)

Necessity is one whose income elasticity is greater than zero and is less than unity. (4.5)

Net investment is gross investment minus depreciation of the existing capital stock. (12.4)

Non-tariff barriers, such as product content requirements, limits the gains from trade. (15.5)

Normal good is one whose demand increases in response to higher incomes. (3.4)

Normative economics offers recommendations that incorporate value judgments. (2.3)

Oligopoly defines an industry with a small number of suppliers. (11.1)

On-the-job training improves human capital through work experience. (13.3)

Opportunity cost of a choice is what must be sacrificed when a choice is made. (1.3)

Ordinal utility assumes that individuals can rank commodity bundles in accordance with the level of satisfaction associated with each bundle. (6.3)

Participation rate: the fraction of the population in the working age group that joins the labour force. (12.2)

Partnership: a business owned jointly by two or more individuals, who share in the profits and are jointly responsible for losses. (7.1)

Patent laws grant inventors a legal monopoly on use for a fixed period of time. (10.7)

Payoff matrix: defines the rewards to each player resulting from particular choices. (11.4)

Percentage change$=$ $[(\text{change in values})/(\text{original value})] \times 100$. (2.1)

Perfect competition: an industry in which many suppliers, producing an identical product, face many buyers, and no one participant can influence the market. (9.1)

Physical capital is the stock of produced goods that are inputs to the production of other goods and services. (12.4)

Portfolio: a combination of assets that is designed to secure an income from investing and to reduce risk. (7.4)

Positive economics studies objective or scientific explanations of how the economy functions. (2.3)

Predatory pricing is a practice that is aimed at driving out competition by artificially reducing the price of one product sold by a supplier. (14.5)

Present value of a stream of future earnings: the sum of each year's earnings divided by one plus the interest rate raised to the appropriate power. (12.4)

Price controls are government rules or laws that inhibit the formation of market-determined prices. (3.7)

Price discrimination involves charging different prices to different consumers in order to increase profit. (10.5)

Price elasticity of demand is measured as the percentage change in quantity demanded, divided by the percentage change in price. (4.1)

Principal or owner: delegates decisions to an agent, or manager. (7.2)

Principal-agent problem: arises when the principal cannot easily monitor the actions of the agent, who therefore may not act in the best interests of the principal. (7.2)

Principle of comparative advantage states that even if one country has an absolute advantage in producing both goods, gains to specialization and trade still materialize, provided the opportunity cost of producing the goods differs between economies. (15.3)

Process innovation refers to new or better production or supply. (10.7)

Product innovation refers to new or better products or services. (10.7)

Production function: a technological relationship that specifies how much output can be produced with specific amounts of inputs. (8.1)

Production possibility frontier (PPF) defines the combination of goods that can be produced using all of the resources available. (1.4)

Productivity of labour is the output of goods and services per worker. (1.6)

Profit maximization is the goal of competitive suppliers – they seek to maximize the difference between revenues and costs. (9.1)

Public goods are non-rivalrous, in that they can be consumed simultaneously by more than one individual; additionally they may have a non-excludability characteristic. (14.1)

Quantity demanded defines the amount purchased at a particular price. (3.2)

Quantity supplied refers to the amount supplied at a particular price. (3.2)

Quota is a quantitative limit on an imported product. (15.5)

Quotas are physical restrictions on output. (3.7)

Reaction functions define the optimal choice of output conditional upon a rival's output choice. (11.5)

Real price: the actual price adjusted by the general (consumer) price level in the economy. (2.1)

Real return on corporate stock: the sum of dividend plus capital gain, adjusted for inflation. (7.1)

Real return: the nominal return minus the rate of inflation. (7.1)

Recession: when output falls below the economy's capacity output. (1.6)

Refusal to deal: an illegal practice where a supplier refuses to sell to a purchaser. (14.5)

Regression line: representation of the average relationship between two variables in a scatter diagram. (2.2)

Rent is the excess remuneration an individual currently receives above the next best alternative. This alternative is the reservation wage. (12.3)

Rent seeking is an activity that uses productive resources to redistribute rather than create output and value. (10.7)

Rental rate: the cost of using capital services. (12.4)

Repeated cross-section data: cross-section data recorded at regular or irregular intervals. (2.1)

Required rental covers the sum of maintenance, depreciation and interest costs. (12.5)

Resale price maintenance is an illegal practice wherein a producer requires sellers to maintain a specified price. (14.5)

Retained earnings are the profits retained by a company for reinvestment and not distributed as dividends. (7.1)

Revenue burden is the amount of tax revenue raised by a tax. (5.4)

Risk: the risk associated with an investment can be measured by the dispersion in possible outcomes. A greater dispersion in outcomes implies more risk. (7.4)

Risk pooling: a means of reducing risk and increasing utility by aggregating or pooling multiple independent risks. (7.4)

Screening is the process of obtaining information by observing differences in behaviour. (13.4)

Shareholders invest in corporations and therefore are the owners. They have limited liability personally if the firm incurs losses. (7.1)

Sharing economy: involves enterprises that are internet based, and that use production resources that have use outside of the marketplace. (14.5)

Short run: a period during which at least one factor of production is fixed. If capital is fixed, then more output is produced by using additional labour. (8.2)

Short-run equilibrium in perfect competition occurs when each firm maximizes profit by producing a quantity where $P = MC$. (9.3)

Short-run supply curve for perfect competitor: the portion of the MC curve above the minimum of the AVC. (9.3)

Short side of the market determines outcomes at prices other than the equilibrium. (3.2)

Shut-down price corresponds to the minimum value of the AVC curve. (9.3)

Signalling is the decision to undertake an action in order to reveal information. (13.4)

Sole proprietor is the single owner of a business and is responsible for all profits and losses. (7.1)

Spending power of a federal government arises when the federal government can influence lower level governments due to its financial rather than constitutional power. (14.2)

Stock is the quantity of an asset at a point in time. (12.4)

Stock option: an option to buy the stock of the company at a future date for a fixed, predetermined price. (7.2)

Strategy: a game plan describing how a player acts, or moves, in each possible situation. (11.4)

Substitute goods: when a price reduction (rise) for a related product reduces (increases) the demand for a primary product, it is a substitute for the primary product. (3.4)

Sunk cost is a fixed cost that has already been incurred and cannot be recovered, even by producing a zero output. (8.5)

Supplier or producer surplus is the excess of market price over the reservation price of the supplier. (5.2)

Supply is the quantity of a good or service that sellers are willing to sell at each possible price, with all other influences on supply remaining unchanged. (3.2)

Supply chain: denotes the numerous sources for intermediate goods used in producing a final product. (15.4)

Supply curve is a graphical expression of the relationship between price and quantity supplied, with other influences remaining unchanged. (3.3)

Tariff is a tax on an imported product that is designed to limit trade in addition to generating tax revenue. It is a barrier to trade. (15.5)

Tax incidence describes how the burden of a tax is shared between buyer and seller. (4.7)

Tax wedge is the difference between the consumer and producer prices. (5.4)

Technological change represents innovation that can reduce the cost of production or bring new products on line. (8.7)

Technological efficiency means that the maximum output is produced with the given set of inputs. (8.1)

Terms of trade define the rate at which goods trade internationally. (15.3)

Theory is a logical view of how things work, and is frequently formulated on the basis of observation. (1.2)

Tied sale: one where the purchaser must agree to purchase a bundle of goods from the one supplier. (14.5)

Time series data: a set of measurements made sequentially at different points in time. (2.1)

Total cost is the sum of fixed cost and variable cost. (8.4)

Total factor productivity: how efficiently the factors of production are combined. (15.4)

Total product is the relationship between total output produced and the number of workers employed, for a given amount of capital. (8.3)

Total utility is a measure of the total satisfaction derived from consuming a given amount of goods and services. (6.2)

Tradable permits are a market-based system aimed at reducing GHGs. (5.7)

Trade subsidy to a domestic manufacturer reduces the domestic cost and limits imports. (15.5)

Transfer earnings are the amount that an individual can earn in the next highest paying alternative job. (12.3)

Two-part tariff: involves an access fee and a per unit of quantity fee. (14.5)

Unemployment rate: the fraction of the labour force actively seeking employment that is not employed. (12.2)

Value of the marginal product is the marginal product multiplied by the price of the good produced. (12.1)

Value of the marginal product of capital is the marginal product of capital multiplied by the price of the output it produces. (12.5)

Variable costs are related to the output produced. (8.4)

Variables: measures that can take on different sizes. (2.1)

Very long run: a period sufficiently long for new technology to develop. (8.2)

Welfare economics assesses how well the economy allocates its scarce resources in accordance with the goals of efficiency and equity. (5.1)

Zero-sum game: an interaction where the gain to one party equals the loss to another party. (1.4)

Glossary: Macroeconomic Terms

AD/AS model: a framework used to explain the behaviour of real output and prices in the national economy.(17.1)

Aggregate demand: *planned aggregate expenditure* on final goods and services at different price levels, all other conditions remaining constant. (17.1)

Aggregate expenditure (AE): *planned expenditure* by business and households. (18.2)

Aggregate expenditure (AE): the sum of *planned induced* and *autonomous* expenditure in the economy. (18.2)

Aggregate production function: the link between labour and capital inputs to production, technology and the output of real GDP. (25.2)

Aggregate supply: the output of final goods and services businesses would produce at different price levels, all other conditions held constant. (17.1)

Appreciation of the national currency: an increase in the value of the currency relative to other national currencies, which results in a fall in the domestic currency price of foreign currencies. (21.4)

Automatic stabilizers: tax and transfer programs that reduce the size of the multiplier and the effects of *transitory* fluctuations in autonomous expenditures on equilibrium GDP. (19.5)

Autonomous expenditure (A): planned expenditure that is not determined by current income. (18.2)

Balance of payments accounts: a record of trade and financial transactions between residents of one country and the rest of the world. (24.1)

Balance of payments: the sum of the balances in current accounts and financial accounts, minus the change in the holdings of official reserves. (24.1)

Balanced budget: revenues are equal to expenditures. (19.3)

Bank of Canada: Canada's central bank. (20.3)

Bank rate: the interest rate the central bank charges on its loans to commercial banks. (22.2)

Bank reserves: cash (legal tender) held by banks to meet possible withdrawals by depositors. (20.1)

Bankers risk: the risk that customers may demand cash for their deposits. (20.3)

Barter exchanges: direct exchanges of goods or services for goods or services without the use of money. (20.1)

Bond coupon: the *annual* fixed money payment paid to a bond holder. (21.1)

Bond price: the *present value* of future payments of interest and principal. (21.1)

Bond: a financial contract that makes one or more fixed money payments at specific dates in the future. (21.1)

Boom: a period of high growth that raises output above normal capacity output. (1.6)

Budget deficit: revenues are less than expenditures. (19.3)

Budget function: the relationship between the budget balance and the level of national income for a specific budget program. (19.3)

Budget surplus: revenues are greater than expenditures. (19.3)

Business cycles: short-term fluctuations of actual real GDP. (17.4)

Capital stock: the buildings, machinery, equipment and software used in producing goods and services. (1.6)

Central bank intervention: purchases or sales of foreign currency intended to manage the exchange rate. (24.3)

Central bank: an institution that conducts monetary policy using its control of monetary base and interest rates. (22.1)

Centralized e-money: a payment card denominated in a national currency. (20.1)

Change in official international reserves: the change in the Government of Canada's foreign currency balances. (24.1)

Circular flow diagrams: show the flows of money payments, real resources, and goods and services between households and businesses. (16.3)

Clearing balances: cash balances used by a bank to settle any differences between its customers' payments and receipts involving other banks. (20.3)

Comparative static analysis: compares an initial equilibrium with a new equilibrium, where the difference is due to a change in one of the conditions behind the initial equilibrium. (3.4)

Complementary goods: when a price reduction (rise) for a related product increases (reduces) demand for a primary product it is a complement for the primary product. (3.3)

Constant returns to scale: equal percentage increases in inputs of labour and capital increase output by the same percentage. (25.3)

Consumer price index (CPI): a measure of the cost of living in any one year to the cost of living in a base year. (16.1)

Consumer price index: the average price level for consumer goods and services. (2.1)

Consumption possibility frontier (CPF): the combination of goods that can be consumed as a result of a give production choice. (1.4)

Convertible currency: a national currency that can be freely exchanged for a different national currency at the prevailing exchange rate. (24.3)

Cost of credit: the cost of financing expenditures by borrowing at market interest rates. (21.4)

Credit easing: the management of the central bank's assets designed to support lending in specific financial markets. (22.4)

Credit money: the debt of a private business or individual. (20.1)

Cross-section data: values for different variables recorded at a point in time. (2.1)

Cryptocurrency: a decentralized e-money with no centralized issuer and not denominated in any national currency. (20.1)

Currency appreciation: a rise in external value of the domestic currency that lowers the domestic currency price of foreign currency. (24.2)

Currency depreciation: a fall in external value of the domestic currency that raises domestic currency price of foreign currency. (24.2)

Current account: a record of trade in goods, services, and transfer payments. (24.1)

Cyclical unemployment: would be eliminated by higher levels of economic activity. (16.1)

Data: recorded values of variables. (2.1)

Decentralized e-money: has no centralized issuer and is not denominated in any national currency. It is a cryptocurrency such as a bitcoin. (20.2)

Deflation: a persistent fall in the general price level. (23.6)

Demand curve: a graphical expression of the relationship between price and quantity demanded with all other influences unchanged. (3.3)

Demand: the quantity of a good or service that buyers wish to purchase at each possible price with all other influences on demand remaining unchanged.(3.2)

Demand for Money (*L*): is demand to hold real money balances measured in terms of purchasing power over goods and services. (21.2)

Deposit multiplier: increase in bank deposits as the result of an increase in monetary base. (20.5)

Depreciation of the national currency: a decline in the value of the currency relative to other national currencies, which results in a rise in the domestic price of foreign currencies. (21.3)

Devaluation (revaluation): a reduction (increase) in the international value of the domestic currency. (24.3)

Discretionary fiscal policy: changes in net tax rates and government expenditure intended to offset *persistent* autonomous expenditure shocks and stabilize aggregate expenditure and output. (19.5)

Disinflation: a persistent fall in the inflation rate. (23.6)

Disposable income (Y_D): national income minus net taxes. (19.2)

E-monies: are multi-purpose pre-paid payment cards, and stored-value cards and gift cards for purchases at specific companies. (20.2)

E-payments: payments made using mobile access to bank accounts eg. by internet ot telephone. (20.2)

Economic equity: the distribution of well-being among members of the economy. (2.2)

Economic growth: an increase in real GDP. (16.1)

Economic growth: the annual percentage change in real GDP or per capita real GDP. (25.1)

Economic recession: national output falls below the economyâĂŹs capacity output. (1.6)

Economy-wide PPF: the set of goods and services combinations that can be produced in the economy when all available productive resources are in use. (1.5)

Effective lower bound (ELB): A bank's policy interest rate cannot be set below a small positive number. (22.4)

Employment rate: percent of the population 15 years of age and over that is employed. (16.1)

Employment: number of adults employed full-time and part-time and self-employed. (16.1)

Endogenous growth: growth determined economic behaviour and policy within the model. (25.5)

Equilibrium GDP: AD = AS, planned expenditure equals current output and provides business revenues that cover current costs including expected profit. (17.1)

Equilibrium price: the price at which quantity demanded equals quantity supplied. (3.2)

Excess demand: amount by which quantity demanded exceeds the quantity supplied at the current price. (3.2)

Excess supply: amount by which quantity supplied exceeds the quantity demanded at the current price. (3.2)

Exchange rate regime: the policy choice that determines how foreign exchange markets operate. (24.3)

Exchange rate target: monetary policy maintains a fixed price for foreign currency in terms of domestic currency. (22.3)

Exogenous variable: a variable with a value determined outside the model. (25.5)

Expenditure-based nominal GDP: sum of the market value of all the final goods and services bought in a given time period, say one year. (16.4)

Fiat money: money the government has declared as legal tender. (20.1)

Final goods and services: goods and services are purchased by the ultimate users. (16.4)

Financial accounts: record of international flows of funds arising from ineternational purchases and sales of real and financial assets. (24.1)

Financial intermediary: a business that specializes in bringing borrowers and lenders together. (20.3)

Financial panic: a loss of confidence in banks and rush to withdraw cash. (20.4)

Financial portfolio: a mixed holding of money and other financial assets, such as bonds and equities, structured to balance expected return and risk. (21.2)

Fintech: the rapidly growing, innovative, applications of computer and internet technologies to facilitate the flow of payments and financial information. (20.1)

Fiscal austerity: cuts in government expenditure and /or increases in taxes aimed at improving the government's budget balance. (19.4)

Fiscal policy: government expenditure and tax changes designed to influence AD. (17.6)

Fixed exchange rate: an exchange rate set by government policy that does not change as a result of changes in market conditions. (24.3)

Flexible exchange rates: foreign exchange rates as determined by the supply and demand in the foreign exchange market without central bank intervention. (24.3)

Foreign exchange rate: the domestic currency price of a unit of foreign currency. (21.3)

Forward guidance: information on the timing of future changes in the central bank's interest rate setting. (22.4)

Frictional unemployment: a result of the time involved in adjusting to changing labour force and employment opportunities. (16.1)

Full employment output Y_p: $Y_p =$ (number of workers at full employment) $\times$ (output per worker). (1.6)

GDP deflator: index of current final output prices relative to base year prices. (16.5)

GDP (Y): the national accounts measure of the sum of actual expenditure in the economy. (18.2)

Government budget: planned government spending and revenue. (19.3)

Government expenditure (G): government spending on currently produced goods and services. (19.2)

Growth accounting: measurement of the contributions of labour, capital, and technology to growth in output. (25.2)

Growth in potential output (Y_P): the result of growth in full employment labour and capital inputs combined with improvements in technology. (17.3)

High (low) frequency data: series with short (long) intervals between observations. (2.1)

Income-based GDP: sum of the factor costs of production of all goods and services plus the net in direct taxes included in market price. (16.4)

Index number: value for a variable, or an average of a set of variables, expressed relative to a given base value. (2.1)

Induced expenditure $(c-m)Y$: planned consumption and imports expenditures that change when income changes. (18.2)

Inferior good: a good for which demand falls in response to higher incomes. (3.4)

Inflation (deflation) rate: the annual percentage increase (decrease) in the level of consumer prices. (2.1)

Inflation rate target: monetary policy objective defined as an announced target inflation rate. (22.3)

Inflation: a persistent rise in the general price level. (16.1)

Inflationary gap: the amount by which actual real GDP exceeds potential GDP. (17.4)

Innovation: the application of new knowledge into production techniques. (25.4)

Interest rate: the current market rate paid to lenders or charged to borrowers. (21.1)

Intermediate inputs: services, materials, and components purchased from other businesses and used in the production of final goods. (16.4)

Invention: the discovery of new knowledge. (25.4)

Investment function, $I=I(i)$: explains the level of planned investment expenditure at each interest rate. (21.4)

Labour force: adults employed plus those not employed but actively looking for work. (16.1)

Legal tender: money that by law must be accepted as a means of payment. (20.1)

Liquidity: the cost, speed, and certainty with which asset values can be converted into cash. (20.3)

Longitudinal data: follow the same units of observation through time. (2.1)

Macroeconomics: the study of the economy as a complete system that determines national output, employment and prices. (1.1)

Marginal product: the change in total output caused by a change of one unit in the input of that factor to production. (25.3)

Marginal propensity to consume ($mpc = c = \Delta C/\Delta Y$): the change in consumption expenditure caused by a change in income. (18.2)

Marginal propensity to import ($mpm = m = \Delta IM/\Delta Y$): the change in imports caused by a change in income. (18.2)

Market demand: the horizontal sum of individual demands. (3.7)

Market value of output≡total expenditure≡market value of factor services≡household income. (16.3)

Means of payment: a commodity or token generally accepted in payment for goods and services or the repayment of debt. (20.1)

Microeconomics: the study of individual behaviour in the context of scarcity. (1.1)

Mixed economy: goods and services are supplied both by private suppliers and government. (1.1)

Model: a formalization of theory that facilitates scientific inquiry. (1.2)

Monetary base (MB): legal tender comprising notes and coins in circulation plus the cash held by the banks. (20.5)

Monetary policy instrument: the monetary variable the central bank manipulates in pursuit of its policy target. (22.3)

Monetary policy: central bank action to control inflation and support economic growth through control of the money supply, interest rates, and exchange rates to change aggregate demand and economic performance. (22.1)

Monetary policy: changes in interest rates and money supply designed to influence AD. (17.6)

Monetary policy rule: central bank interest rate setting based on inflation rate and output gap targets. (22.2)

Monetary policy indicators: variables that provide information about the stimulus or restraint coming from the central bank's policy. (22.5)

Money supply target: a central bank adjusts interest rates and the monetary base to control the nominal money supply, or the rate of growth of the nominal money supply. (22.3)

Money supply: notes and coin in circulation outside banks plus bank deposits. (20.2, 20.5)

Money supply: the means of payment in the economy, namely currency (notes and coin) in circulation outside the banks and bank deposits. (20.1)

Moral suasion: a central bank persuades and encourages banks to follow its policy initiatives and guidance. (22.4)

Multiplier ($\Delta Y/\Delta A$): the ratio of the change in equilibrium income Y to the change in autonomous expenditure A that caused it. (18.4)

NAIRU: the 'non-accelerating inflation rate of unemployment' that corresponds to N_F at Y_P. (23.2)

National Accounts: a framework for measuring national income, GDP, based on expenditures on goods and services and incomes earned. (16.3)

Natural unemployment rate: the unemployment rate at "full employment". (16.1)

Natural unemployment rate: the unemployment rate that corresponds to potential GDP. (17.2)

Net interest income: the excess of loan interest earned over deposit interest paid. (20.3)

Net taxes: taxes on incomes minus transfer payments. (19.2)

Neutral rate of interest: the interest rate that, in the central bank's judgment, is consistent with actual real GDP equal to potential real GDP inflation at its target. (22.4)

Net Exports: balance of trade in goods and services in the current account of the balance of payments. (24.1)

Nominal exchange rate (er): the domestic currency price of a unit of foreign currency. (24.1)

Nominal GDP: the output of final goods and services, the money incomes generated by the production of that output, and expenditure on the sale of that output measured at current prices in a specific time period. (16.4)

Normal good: a good for which demand increases in response to higher incomes. (3.4)

Normative economics: offers recommendations that incorporate value judgments. (2.2)

Official exchange reserves: government foreign currency holdings managed by the central bank. (24.3)

Official financing: balance of payments entry that records changes in government holdings of foreign currencies. (24.3)

Open market operation: central bank purchases or sales of government securities in the open financial market. (22.2)

Opportunity cost: what must be sacrificed when a choice is made. (1.3)

Output gap: the difference between actual output and potential output. (17.4)

Overnight rate: the interest rate large financial institutions receive or pay on loans from one day until the next. (22.3)

Participation rate: percent of the population that is either working or unemployed. (16.1)

Per capita real GDP: real GDP per person. (16.6)

Percentage change: $[(\text{change in value})/(\text{original value})] \times 100$. (2.1)

Perfect capital mobility: when very small differences in expected returns cause very large international flows of funds. (24.1)

Positive economics: studies objective or scientific explanations of how the economy functions. (2.2)

Potential output: the real GDP the economy can produce on a sustained basis with current labour force, capital and technology without generating inflationary pressure on prices. (17.2, 23.2)

Present value is the *discounted* value of future payments. (21.1)

Price controls: government rules or laws that inhibit the formation of market determined prices. (3.7)

Price index: a measure of the price level in one year compared with prices in a base year. (16.1)

Price level: a measure of the average prices of all goods and services produced in the economy. (16.1)

Price of a marketable bond: the current price at which the bond trades in the bond market. (21.1)

Primary budget balance: the difference between a government's net tax revenues at Y_p and its program expenditures. (23.5)

Prime lending rate: the base for setting the interest rates charged by banks on loans and lines of credit. (22.3)

Production function: outputs determined by technology and inputs of labour and capital. (23.2)

Production possibility frontier (PPF): the combination of goods that can be produced using all resources available. (1.4)

Productivity of labour: the output of goods and services per worker. (1.6)

Productivity: output per unit of input. (23.2)

Public debt (PD): the outstanding stock of government bonds issued to finance government past budget deficits. (19.6)

Public debt ratio (PD/Y): the ratio of outstanding government debt to GDP. (19.6, 23.5)

Quantitative easing: central bank purchases of financial assets to increase its asset holdings and the monetary base. (22.4, 23.6)

Quantity demanded: the amount purchased at a particular price. (3.2)

Quantity supplied: the amount supplied at a particular price. (3.2)

Quotas: are physical restrictions on output. (3.7)

Rate of economic growth: the annual percentage change in real GDP. (16.1)

Real exchange rate: the relative price of goods and services from different countries measured in a common currency. (24.1)

Real GDP: the quantity of final goods and services produced by the economy in a specified time period. (16.1)

Real money supply (M/P): the nominal money supply M divided by the price level P. (21.2)

Real price: the actual price adjusted to a base year value using the consumer price index, the CPI. (2.1)

Recession: decline in economic activity, often defined as two consecutive quarters of negative growth in real GDP. (16.2)

Recessionary gap: amount by which actual real GDP (Y) is less than potential GDP (Y_p). (17.4)

Regression line: representation of the average relationship between two variables in a scatter diagram. (2.2)

Repeated cross-section data: cross-section data recorded at regular or irregular intervals. (2.1)

Required reserve ratio: a legal minimum ratio of cash reserves to deposits. (22.2)

Reserve ratio (rr): the ratio of cash reserves to deposit liabilities held by banks. (20.4)

Short run: a time frame in which factor prices, supplies of factors of production, and technology are fixed by assumption. (17.1)

Short run production function: the relationship between the inflation rate and the level of real GDP. (23.2)

Short-run equilibrium output: Aggregate expenditure and current output are equal ($Y = \text{AE}$). (18.3)

Solow residual: the growth in real GDP or per capita real GDP not caused by growth in factor inputs, but attributed to improved technology. (25.2)

SPRA: A Bank of Canada purchase of securities one day combined with an agreed resale of the securities the next day. (22.3)

SRA: A Bank of Canada sale of securities one day combined with an agreed repurchase of the securities the next day. (22.3)

Standard of deferred payments: the units in which future financial obligations are measured. (20.1)

Store of value: an asset that carries purchasing power forward in time for future purchases. (20.1)

Structural budget balance (SBB): the government budget balance at potential output. (19.4)

Structural primary government balance ($SPBB = tY_P - G$): the difference between net tax revenue at Y_P and government program expenditure. It excludes interest payments on the public debt and the effect of output gaps. (23.5)

Structural unemployment: caused by changes in economic structure relative to labour characteristics. (16.1)

Substitute goods: when a price reduction (rise) for a related product reduces (increases) demand for the primary product, then it is a substitute for the primary product. (3.4)

Supply curve: a graphical expression of the relationship between price and quantity supplied with all other influences unchanged. (3.3)

Supply: the quantities of a good or service that sellers are willing to sell at each possible price with all other influences on supply reaming unchanged. (3.2)

Sustained growth in per capita real GDP: improvements in technology overcome the diminishing returns to increases in the capital to labour ratio. (25.4)

Taylor rule: explains the central bank's policy interest rate settings based on the economy's performance relative to the bank's inflation and output targets. (22.4)

Theory: a logical view of how things work. Frequently formulated on the basis of observation. (1.2)

Time-series: a set of measurements made sequentially at different points in time. (2.1)

Token money: convertible claims on commodity money. (20.1)

Total factor productivity (TFP): measures by the economy's real output relative to the combined inputs of labour and capital, the total factor inputs to production. (25.2)

Transmission mechanism: links money, interest rates, and exchange rates through financial markets to output and employment and prices. (21.4)

Unemployment rate is the number of unemployed persons as a percentage of the labour force. (16.1)

Unemployment: number of adults not working but actively looking for work. (16.1)

Unit of account: the standard in which prices are quoted and accounts are kept. (20.1)

Unplanned changes in business inventories: indicators of disequilibrium between planned and actual expenditures – incentives for businesses to adjust levels of employment and output (Y). (18.3)

Value added: the difference between the market value of the output of the business and the cost of inputs purchased from other businesses. (16.4)

Variables: measures that can take on different values. (2.1)

Very long run: the time required for changes to occur in the stock of capital, the size of the labour force, and the technology of production. (25.2)

Wealth effect: the change in expenditure caused by a change in real wealth. (21.4)

Yield on a bond: the return to a bond holder expressed as an annual percentage. (21.1)

Zero-sum game: an interaction where the gain to one party equals the loss to another party. (1.4)

Solutions To Exercises

Chapter 1 Solutions

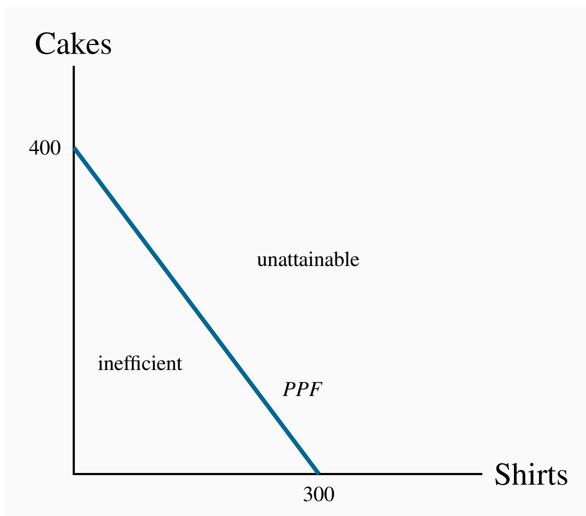
Exercise 1.1

(a) If all 100 workers make cakes their output is $100 \times 4 = 400$.

(b) If all workers make shirts their output is $100 \times 3 = 300$.

(c) The diagram shows the PPF for this economy.

(d) As illustrated in the diagram.



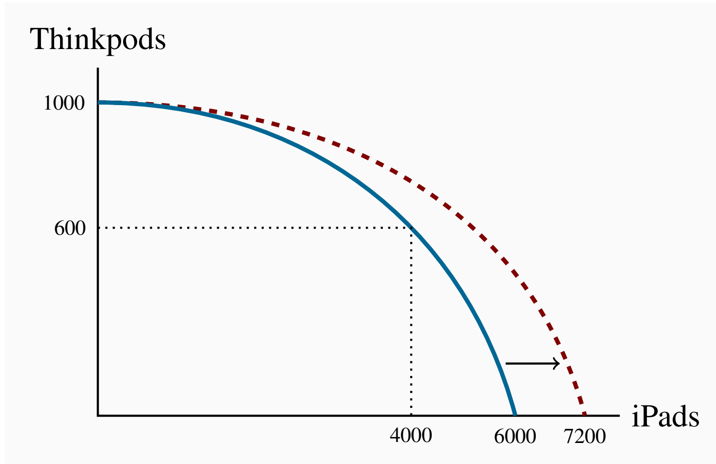
Exercise 1.2

(a) The PPF is curved outwards with intercepts of 1000 on the Thinkpod axis and 6000 on the iPad axis. Each point on the PPF shows one combination of outputs.

(b) Different.

(c) 400 X.

(d) The new PPF in the diagram has the same Thinkpod intercept, 1000, but a new iPad intercept of 7200.

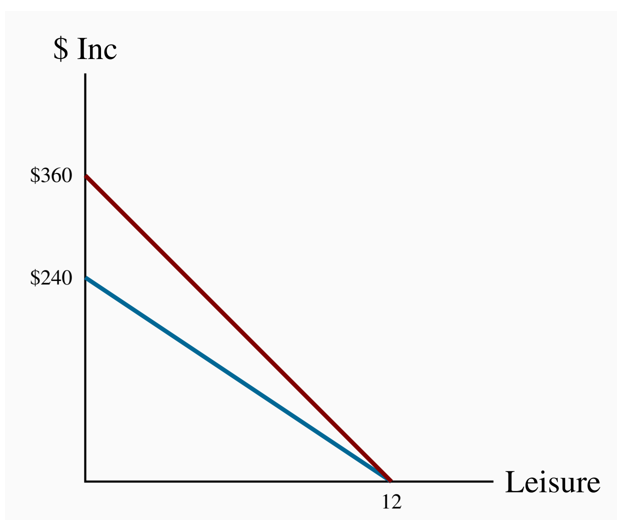


Exercise 1.3

By examining the opportunity cost in the region where the combinations are defined, and by assuming a linear trade-off between each set of combinations, it can be seen that the first combination in the table is feasible, but not the second combination.

Exercise 1.4

(a) \$20.
(b) \$30.
(c) See lower curve on diagram.
(d) See upper curve on diagram.



Exercise 1.5

(a) Louis has an advantage in cutting the grass while Carrie Anne should wash cars.

(b) If they each work a twelve-hour day, between them they can cut 12 lawns and wash 24 cars.

Exercise 1.6

(a) Carrie Anne's lawn intercept is now 12 rather than 8.

(b) Yes, specialization still matters because C.A. is more efficient at cars.

(c) The new coordinates will be 39 on the vertical axis, 24 on the horizontal axis and the kink point is the same.

Exercise 1.7

(a) 220 cakes requires 55 workers, the remaining 45 workers can produce 135 shirts. Hence this combination lies inside the *PPF* described in Exercise 1.1.

(b) 98 workers.

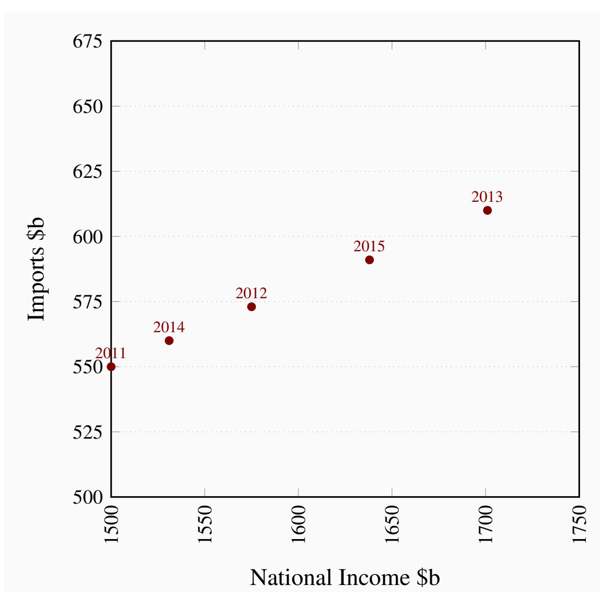
(c) 2%.

Chapter 2 Solutions

Exercise 2.1

(a) Imports and national income are positively related.

(b)



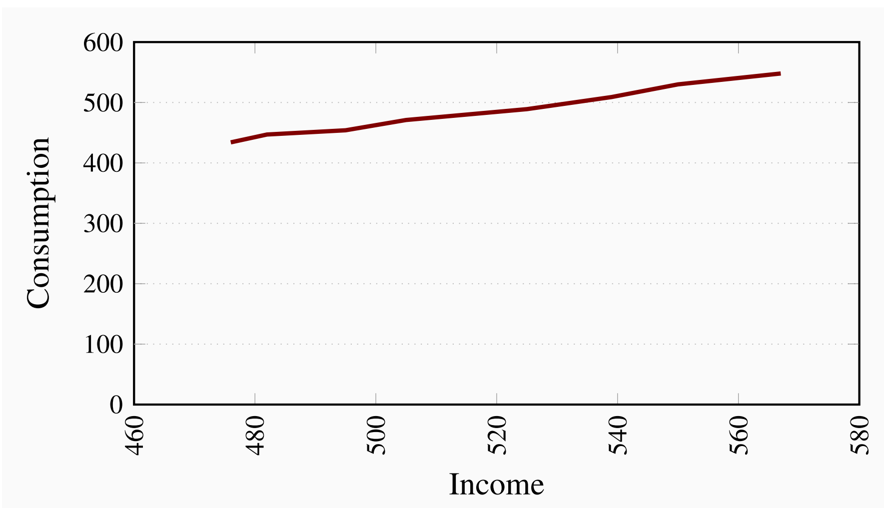
Exercise 2.2

For (b) the answer is 28.6%, for (c) the answer is 3.8%, and for (d) the answers are 3.3% and -2.0%.

Year	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
Index	100.0	104.8	108.6	114.3	119.0	119.0	126.7	128.6	138.1	141.9

Exercise 2.3

The scatter diagram plots observed combinations of income and consumption as follows. For parts (c) and (d): the variables are positively related and the causation runs from income to consumption.



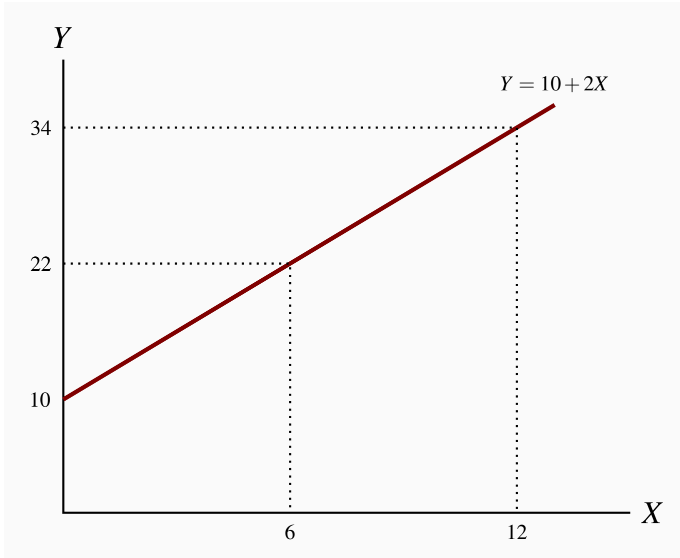
Exercise 2.4

The percentage changes in income are:

Pct Inc	1.3	2.7	2.0	4.0	2.7	2.0	3.1
Pct Con	3.0	1.6	3.7	3.8	4.1	4.1	3.4

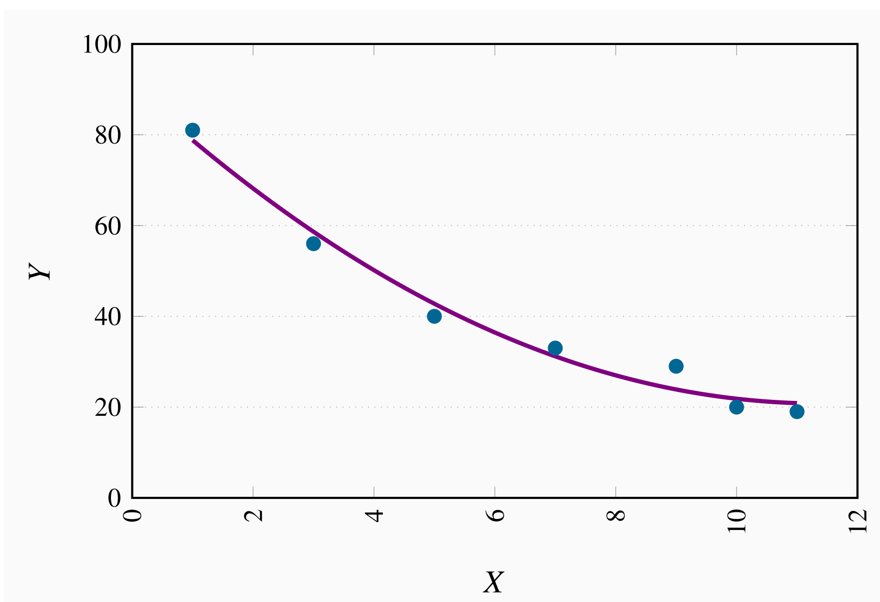
Exercise 2.5

The relationship given by the equation $Y = 10 + 2X$ when plotted has an intercept on the vertical (Y) axis of 10 and the slope of the line is 2. The maximum value of Y (where X is 12) is 34.



X	0	1	2	3	4	5	6	7	8	9	10	11	12
Y	10	12	14	16	18	20	22	24	26	28	30	32	34

Exercise 2.6



(a) The relationship is negative.

(b) The relationship is non-linear.

Chapter 3 Solutions

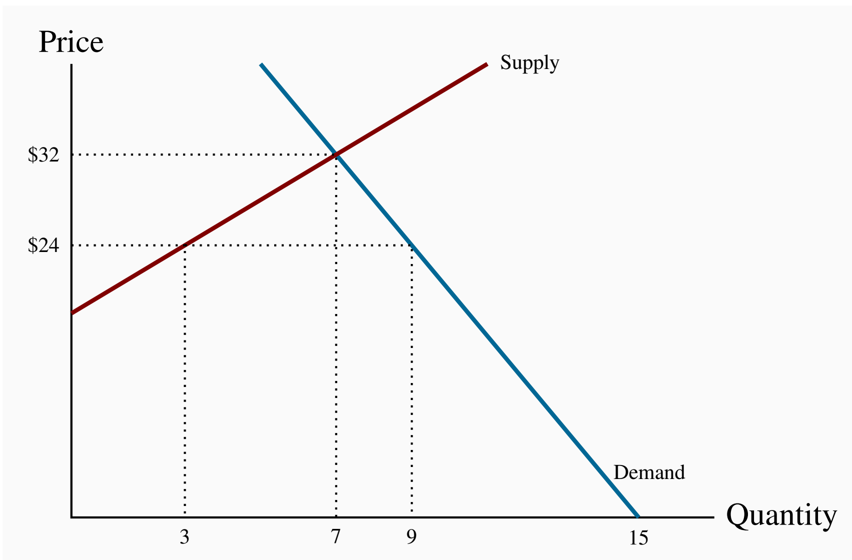
Exercise 3.1

(a) The diagram shows the supply and demand curves from the data in the table. These curves intersect at the equilibrium price \$32 and the equilibrium quantity 7.

(b) Excess demand is 6 and excess supply is 3.

(c) With excess demand the price is bid up, with excess supply the price is pushed down.

(d) Equate supply P to demand: $18+2Q=60-4Q$, implying $6Q=42$, which is $Q=7$. Hence $P=32$.



Exercise 3.2

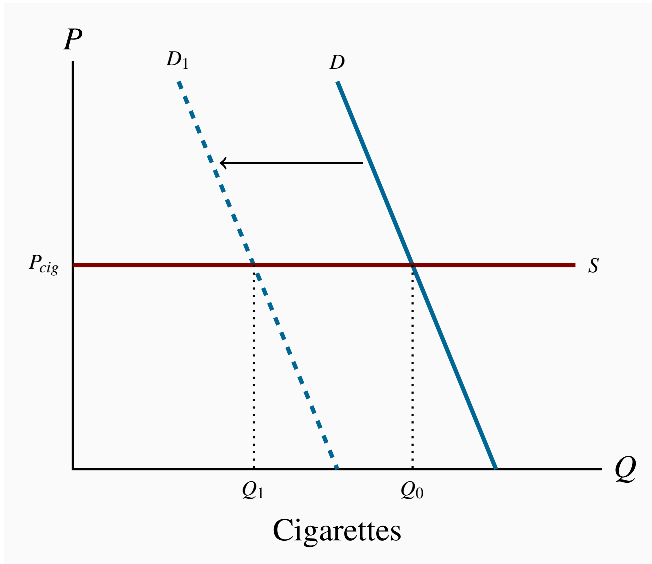
(a) Demand curve facing *Air Canada* shifts left and down. The price of the substitute *Via Rail* has fallen and reduced the quantity of air transport services demanded at any price.

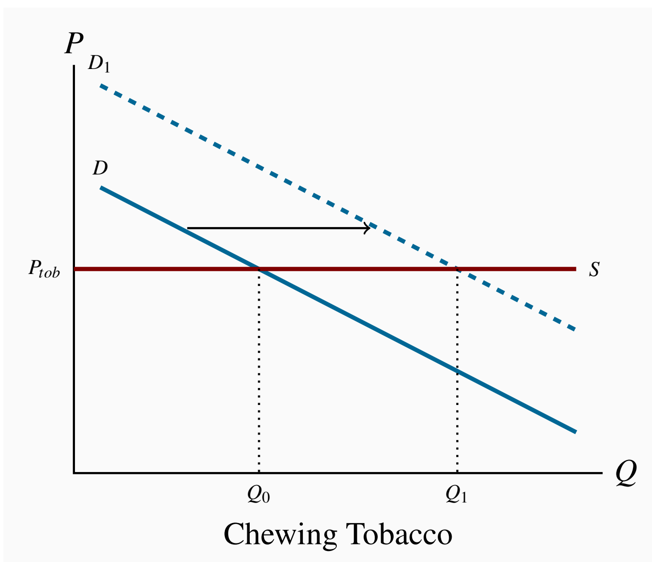
(b) Demand curve facing *Air Canada* shifts left and down. The substitute car travel has improved in quality and perhaps declined in cost.

(c) Demand curve facing *Air Canada* shifts left and down. A new budget air carrier is another substitute for *Air Canada* that will divide the market for air transport.

Exercise 3.3

The market diagrams are drawn on the assumption that each product can be purchased for a given price, the supply curve in each market segment is horizontal. A downward sloping demand should characterize each market. If the cigarette market is 'quashed' the demand in the market for chewing tobacco, a substitute, should shift outward, leading to higher consumption at the same price.

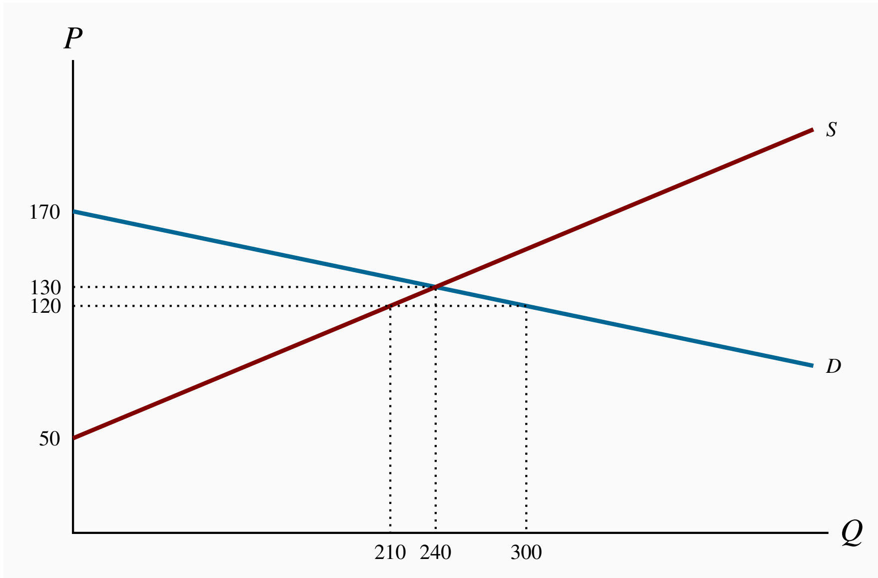




Exercise 3.4

(a) The diagram shows that equilibrium quantity is 240, equilibrium price is \$130, which are the values obtained from equating supply and demand.

(b) At a price of \$120 the quantity demanded is 300 and the quantity supplied 210. Excess demand is therefore 90.



Exercise 3.5

(a) At a price of \$140 quantity demanded is 180 and quantity supplied is 270; excess supply is therefore 90.

(b) Total quotas of 180 will maintain a price of \$140. This is obtained by substituting the price of \$140 into the demand curve and solving for Q.

Exercise 3.6

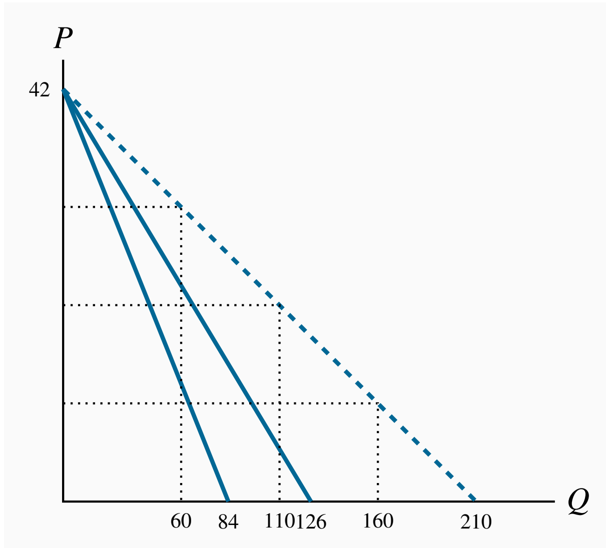
It must buy 90 units at a cost of \$140 each. Hence it incurs a loss on each unit of \$60, making for a total loss of \$5,400.

Exercise 3.7

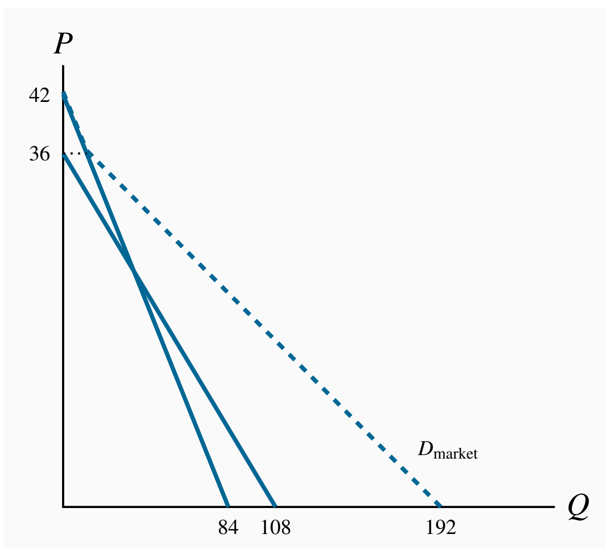
(a) The quantity axis intercepts are 84 and 126.

(b) The quantities demanded are 160, 110 and 60 respectively, on the market demand curve in the diagram. These values are obtained by solving the quantity demanded in each demand equation for a given price and summing the quantities.

(c) The equation for the market demand curve is: $Q = 210 - 5P$.



Exercise 3.8

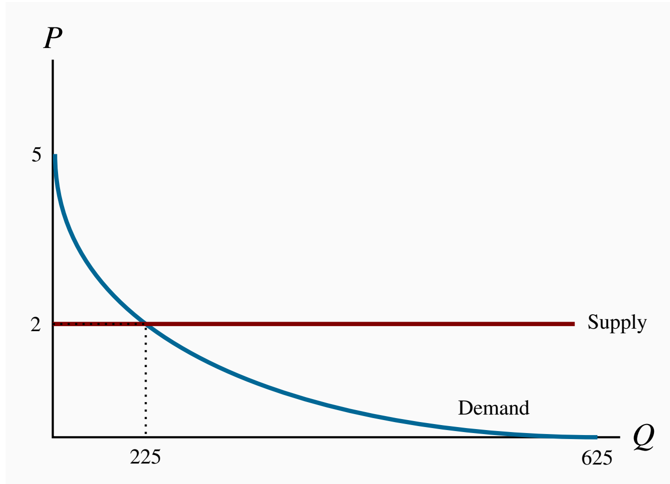


Exercise 3.9

(a) The demand curve is plotted below.

(b) The supply function is horizontal as plotted.

(c) Equilibrium quantity traded at $P = 2$ is $Q = 225$.



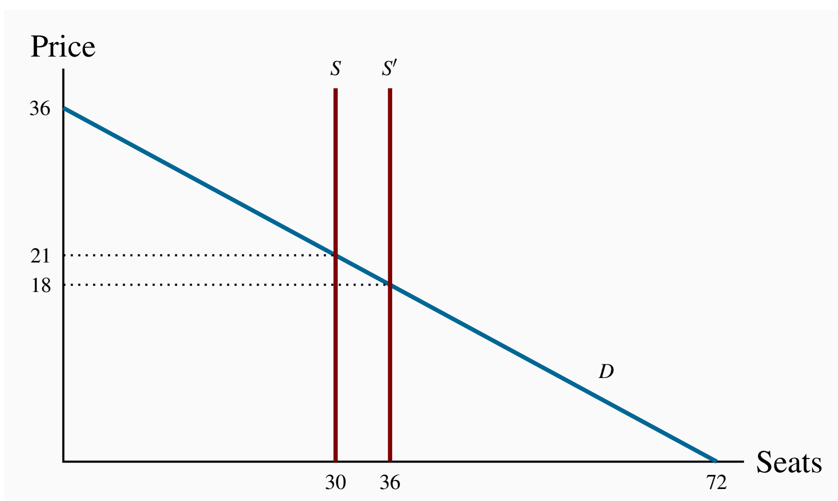
Exercise 3.10

(a) See the diagram below.

(b) The equilibrium admission price is $P = \$21$, $TR = \$630$.

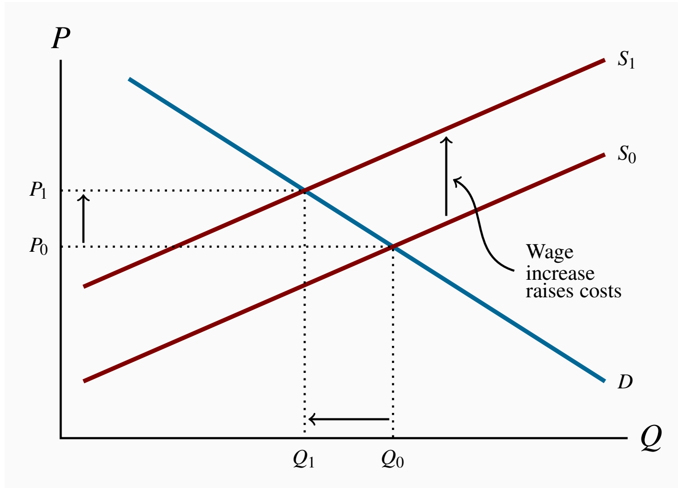
(c) The equilibrium price would now become \$18 and $TR = \$648$. Yes.

(d) The answer is no, because total revenue falls.



Exercise 3.11

Wages are a cost of bringing lettuce to market. In the market diagram the supply curve for lettuce shifts upwards to reflect the increased costs. If demand is unchanged the price of lettuce rises from P_0 to P_1 and the quantity demanded falls from Q_0 to Q_1.



Chapter 4 Solutions

Exercise 4.1

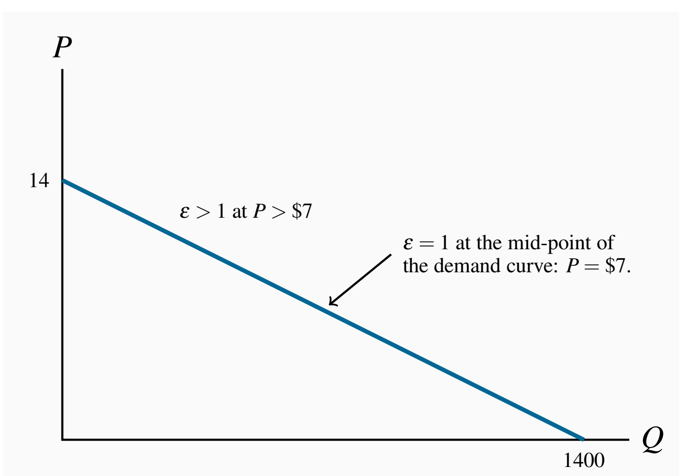
(a) The intercepts for this straight line demand curve are $P = \$14$, $Q = 1400$.

(b) Total revenue is this product of price times quantity. Compute it!

(c) At $P = \$7$, total revenue is \$4,900.

(d) Elasticities, in descending order, are 0.22, 0.33, 0.47, 0.65, 0.87, 1.15.

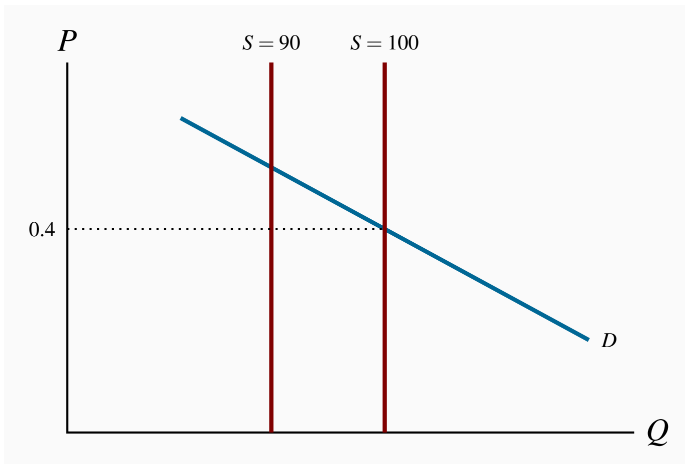
(e) Elasticity becomes greater than one in magnitude at one point where total revenue is maximized.



Exercise 4.2

(a) The supply curve is vertical at a quantity of 100.

(b) We are told $-0.5 = \%\Delta Q/\%\Delta P$. The percentage change of quantity is $-10/95$; therefore the percentage change in price must be: $\%\Delta P = -(10/95)/-0.5 = 20/95 = 21\%$. The new price is therefore $0.4 \times 1.21 = 0.48$.

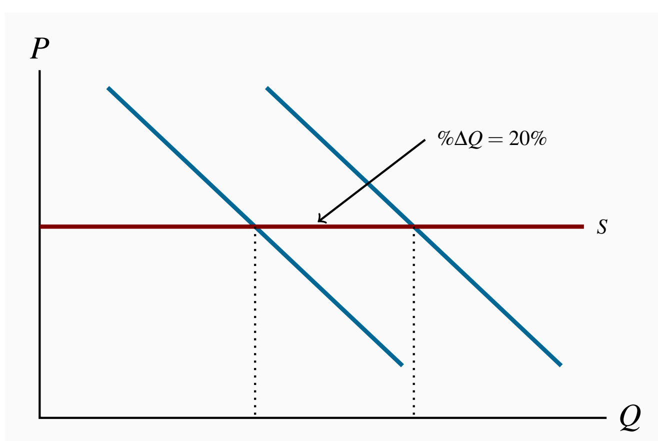


Exercise 4.3

(a) Since the price is fixed the supply curve is horizontal. See figure below.

(b) You cannot estimate a demand elasticity value since there has been no price change.

(c) Here the (adjoining) horizontal supply curve shifts upwards by 60%. If enrolment has increased the demand curve must also have shifted upwards. Draw an additional supply curve representing a 60% upward shift, and find an intersection between the new demand and new supply such that the percentage increase in quantity is 15% (this diagram is not included here).

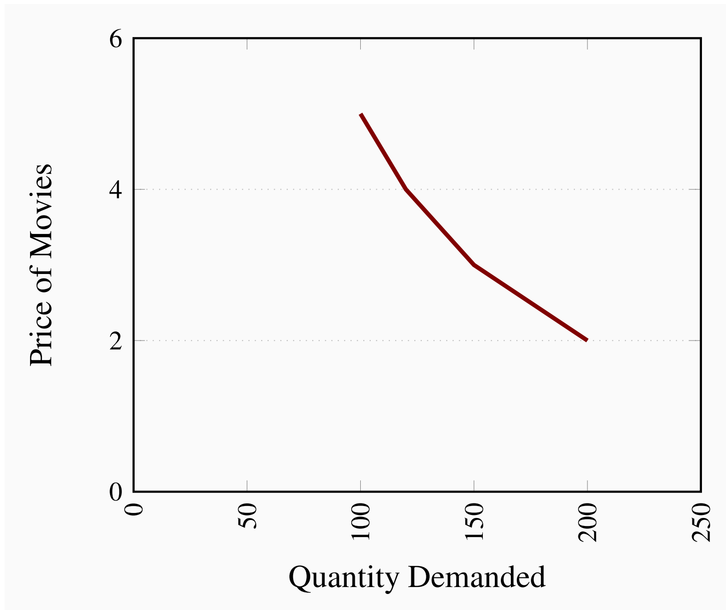


Exercise 4.4

(a) The demand curve is nonlinear.

(b) Total revenue is price times quantity.

(c) Elasticity values are 0.71, 0.78, and 0.82 respectively.



Exercise 4.5

(a) There has been a 20% increase in price. Feeding this into the elasticity formula yields $-0.6 = \%\Delta Q/20\%$. Hence the percentage change (reduction) in quantity is 12%.

(b) Following the same reasoning as in part (a) the result is 24%.

(c) In the short run revenue rises since demand is inelastic (less than one in absolute value); in the long run it falls since demand is elastic (greater than one in absolute value).

Exercise 4.6

(a) It is elastic for magazines and inelastic for CDs and Cappuccinos.

(b) A reduction in magazines purchased and an increase in cappuccinos purchased. Magazines are complements and cappuccinos are substitutes for CDs.

(c) The demand curve for magazines shifts down in response to an increase in the price of CDs and it increases in response to an increase in the price of cappuccinos.

Exercise 4.7

(a) Reduce the price, because the elasticity is greater than one.

(b) Yes, it would reduce train ridership because the positive cross-price elasticity indicates that these goods are substitutes.

Exercise 4.8

(a) Plot the scatter.

(b) The scatter is a positively sloping group of points indicating a positive relationship.

(c) The elasticities estimated at mid values are 1.0, 0.64, 0.47, 0.52, 0.29 and 0.32. For example: the first pair of points yields a $\%\Delta P = 5/12.5$ and $\%\Delta Q = 8{,}000/20{,}000$. Hence, $\%\Delta Q/\%\Delta P = (8{,}000/20{,}000)/(5/12.5) = 1.0$.

(d) They are normal goods because the income elasticity is positive.

Exercise 4.9

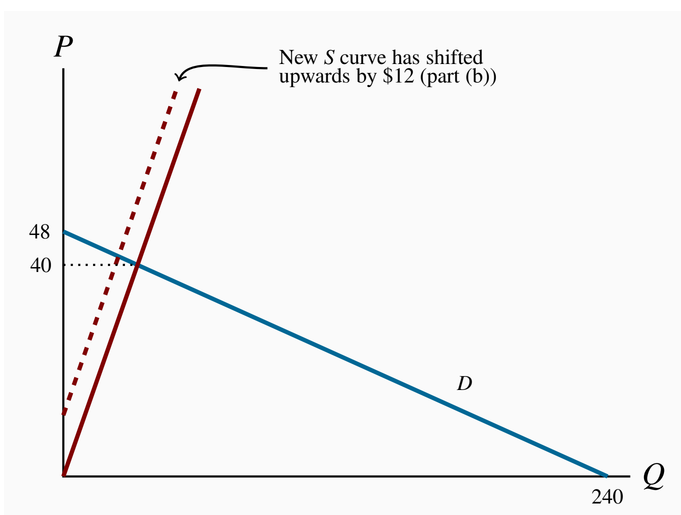
(a) The price intercept for the demand curve is 48 and the quantity intercept is 240. The supply curve goes through the origin with a slope of 1. The equilibrium price is \$40 and the equilibrium quantity is 40.

(b) The supply curve shifts upwards everywhere by \$12.

(c) The price will increase to \$42 and the quantity declines to 30.

(d) The curve still goes through the origin but with a slope of 1.3 rather than 1.0 (not illustrated in the figure).

(e) The equilibrium quantity is $Q = 32$; corresponding price is $P = 32 \times 1.3 = 41.6$.

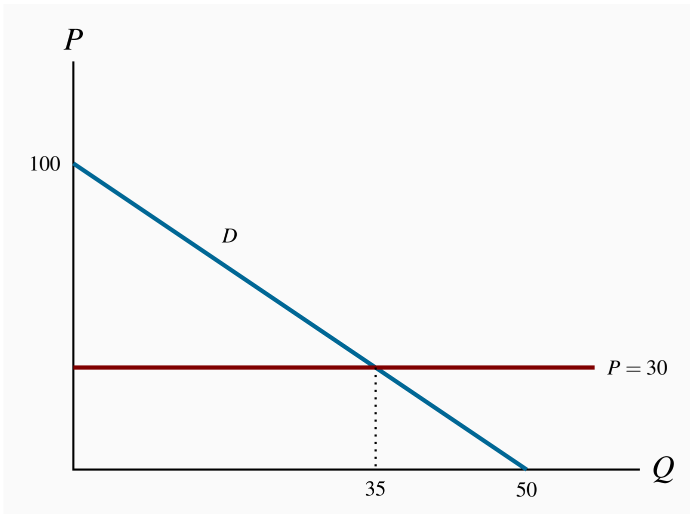


Exercise 4.10

(a) The demand curve has a price intercept of 100 and a quantity intercept of 50. The supply curve is horizontal at a price of \$30. The equilibrium quantity is 35 units at this price.

(b) The new, tax-inclusive, supply curve is horizontal at $P = \$40$ (not illustrated in the figure). The equilibrium price is \$40 and the equilibrium quantity becomes 30. With 30 units sold, each generating a tax of \$10, total tax revenue is \$300.

(c) Since the equilibrium is on the lower half of a linear demand curve the demand is inelastic.

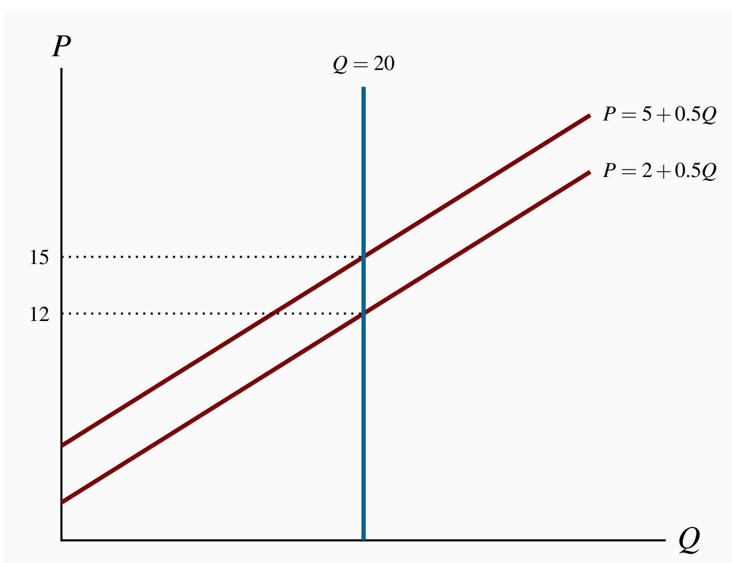


Exercise 4.11

(a) The supply and demand curves are illustrated below.

(b) Solving the demand equations for $Q = 20$ yields prices of \$12 and \$15 respectively.

(c) The consumer bears the entire tax burden.



Chapter 5 Solutions

Exercise 5.1

(a) The step functions are similar to those in Figure 5.1. In ascending order, Margaret is the first supplier, Liam the second, etc. You must also order the demanders in descending order.

(b) Two: Margaret and Liam will supply, while Jones and Lafleur will purchase. The third highest demander (Murray) is willing to pay \$6, while the third supplier is willing to supply only if the price is \$9. Hence there is no third unit supplied.

(c) The equilibrium price will lie in the range \$7.0-\$7.5. So let us say it is \$7. The consumer surplus of each buyer is therefore \$1 and \$0.5. The supplier surpluses are zero and \$2.

(d) Two driveways will still be cleared. The highest value buyers are now willing to pay \$12 and \$8. The third highest value buyer is willing to pay \$7.0. But on the supply side the third supplier still supplies only if he gets \$9. Therefore two units will be supplied. If the price remains at \$7 (it could fall in the range between \$7 and \$8) the consumer surpluses are now \$5 and \$1, and the supplier surpluses remain the same.

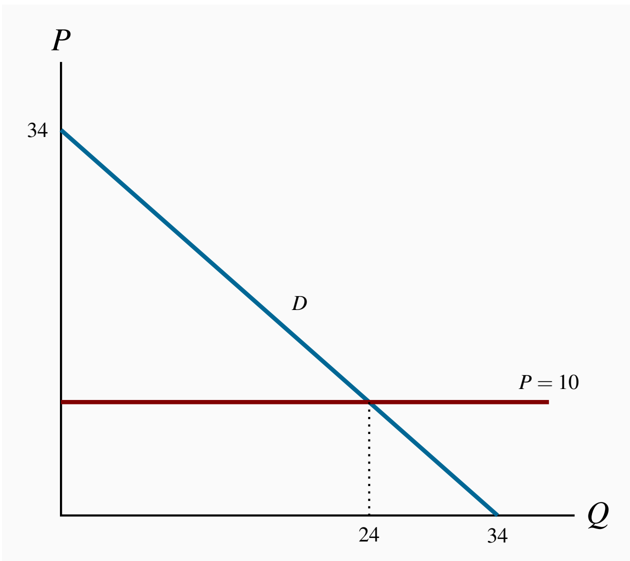
Exercise 5.2

(a) The supply curve is horizontal at a price of \$10. The demand curve price intercept is \$34 and the quantity intercept is 34. The equilibrium quantity is 24.

(b) The new supply curve is $P = 12$, yielding an equilibrium $Q = 22$.

(c) Tax revenue is \$44: each of the 22 units sold yields \$2.

(d) The DWL is \$2.

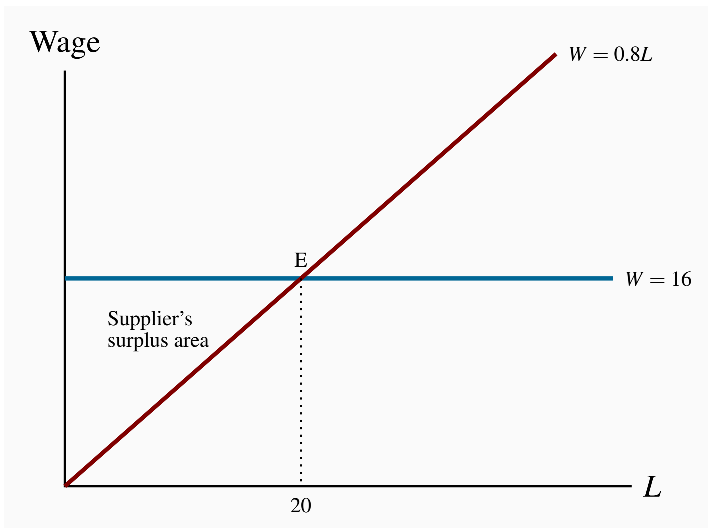


Exercise 5.3

(a) The supply curve goes through the origin and the demand curve is horizontal at $W = \$16$ – see diagram below.

(b) The equilibrium amount of labour supplied is 20 units. The supplier surplus is the area above the supply curve below the equilibrium price= \$160.

(c) At a net wage of \$12, labour supplied falls to 15. The downward shift in the wage reduces the quantity supplied. The new supplier surplus is the triangular area bounded by $W = 12$ and $L = 15$. Its value is therefore \$90.



Exercise 5.4

(a) The demand curve shifts inwards.

(b) Yes, because consumers previously did not have full information about the product.

Exercise 5.5

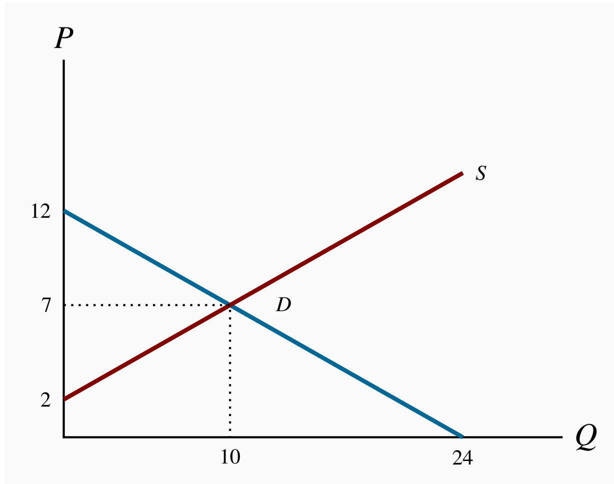
(a) Yes.

(b) Yes, because the congestion effect is not incorporated into the price of driving.

Exercise 5.6

(a) The free market equilibrium is obtained by equating demand and private-cost supply curves: $Q = 10$, $P = \$7$.

(b) The true supply is above the private supply. The social optimum involves a lower output and higher price.



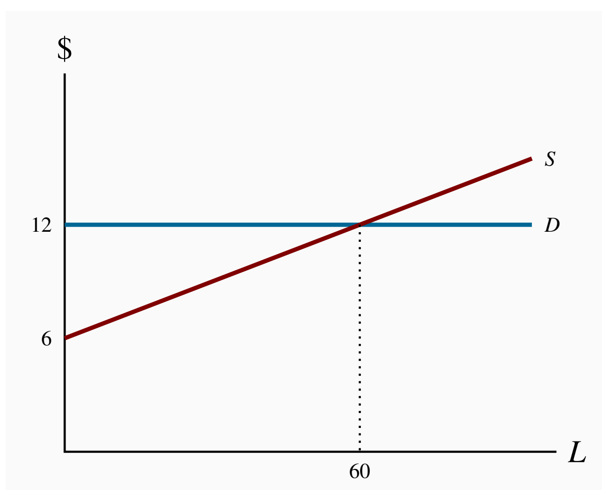
Exercise 5.7

(a) The supply curve is horizontal at $P = \$1$. The demand curve has a price intercept of 5 and a quantity intercept of 1000. The equilibrium quantity is 800.

(b) The socially optimal quantity is obtained by recognizing that the social cost is \$1.25 rather than \$1.0. The supply curve that includes social costs is now horizontal at $P = \$1.25$.

(c) $Q^* = 750$.

Exercise 5.8



(a) The demand curve is horizontal at $P = \$12$. The supply curve slopes upwards with a price intercept of \$6. Equilibrium is $L = 60$.

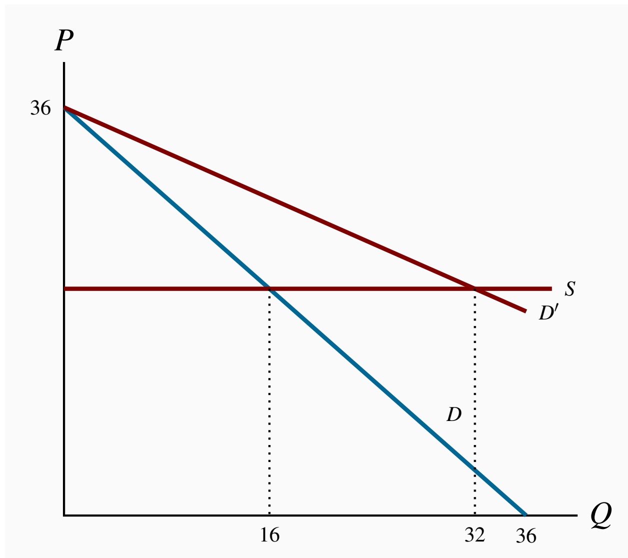
(b) Surplus is the area beneath the demand curve above the supply curve= \$180.

Exercise 5.9

(a) See the diagram below.

(b) The market solution is obtained by equating the market demand and supply. This yields $Q = 16$ and $P = \$20$.

(c) The socially optimal amount takes account of the fact that there are positive externalities. The demand curve that reflects these externalities is above the private demand curve. Hence the socially optimal equilibrium is at a greater output $Q = 32$.



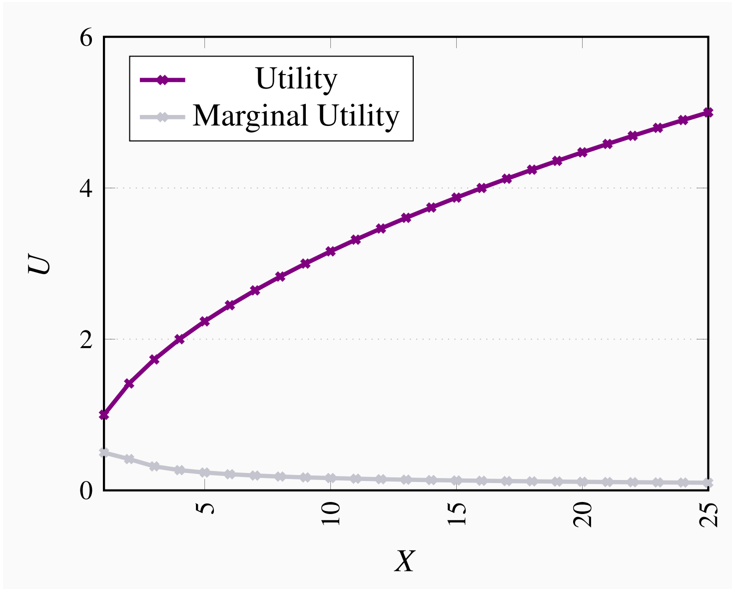
CHAPTER 6 SOLUTIONS

Exercise 6.1

Since the additional utility per dollar spent on another unit of either activity is the same (1.2 units), he should be indifferent as to where he spends it. However, if he gets an income increase that is sufficient to cover the purchase of one whole unit of the goods then snowboarding yields the highest MU per dollar spent.

Exercise 6.2

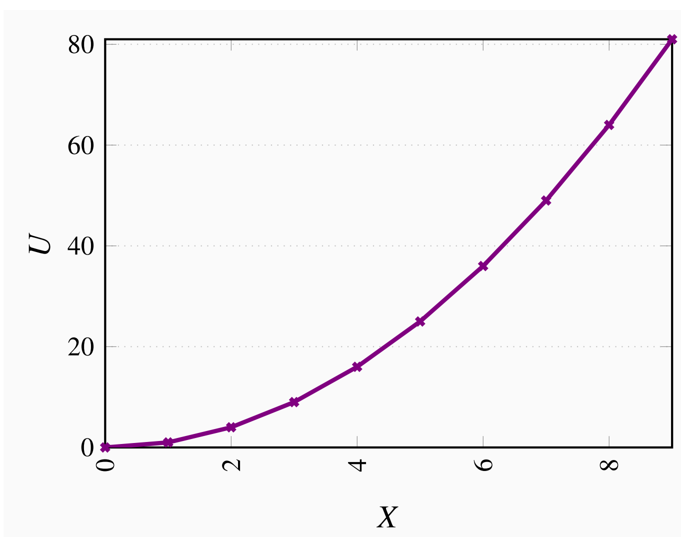
The utility and marginal utility curves are given below.



Exercise 6.3

(a) See the diagram below.

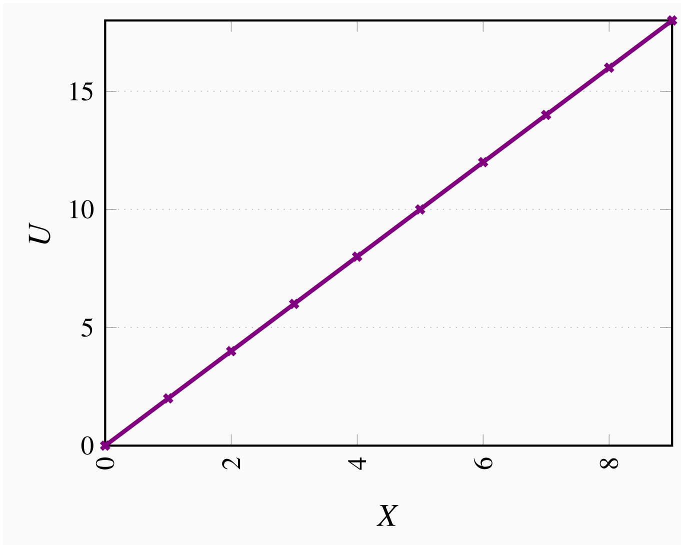
(b) Because the MU increases with each unit consumed.



Exercise 6.4

(a) See the diagram below.

(b) The MU is constant, rather than diminishing.

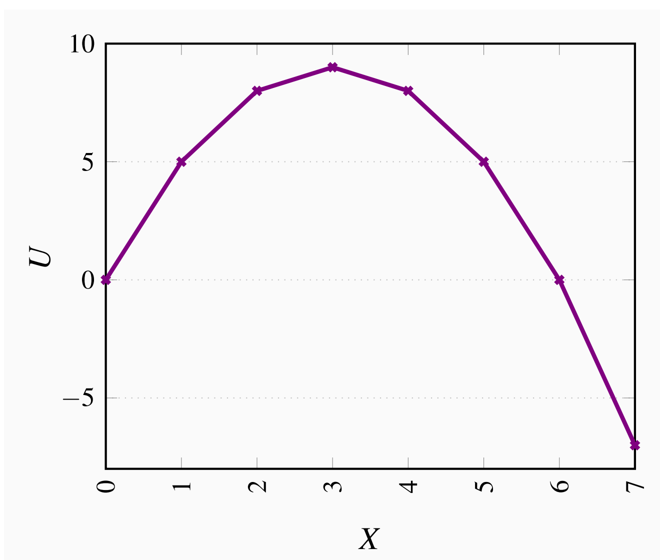


Exercise 6.5

(a) See the diagram below.

(b) $X = 3$.

(c) $X = 6$.



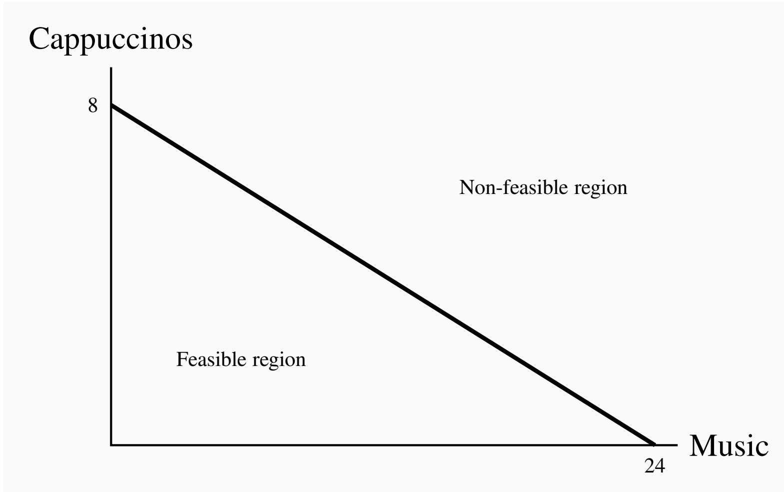
Exercise 6.6

(a) See the diagram below.

(b) The slope is $-8/24 = -1/3$. Opportunity cost of 1 cappuccino is 3 ‘tunes’.

(c) Yes, yes, yes.

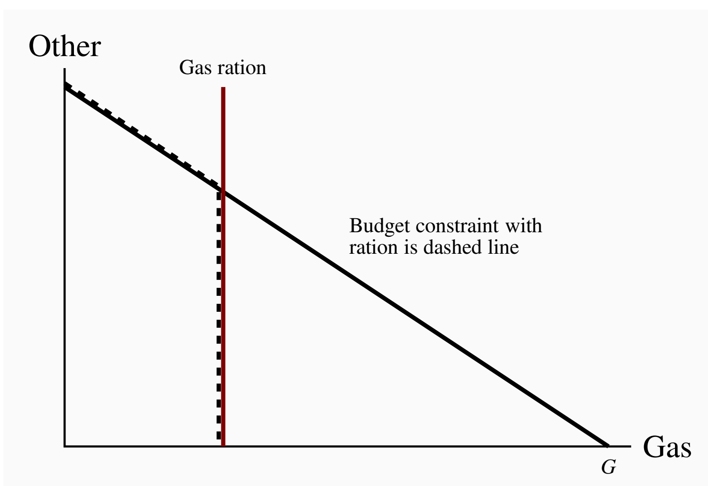
(d) The first two lie inside, the third lies on the budget line.



Exercise 6.7

(a) Let G be the initial intercept on the gasoline axis, then $1/2G$ is the new intercept.

(b) A vertical line at a point less than $1/2G$ reduces the feasible set to the area bounded by the new budget constraint (dashed line) and the vertical line GQ.

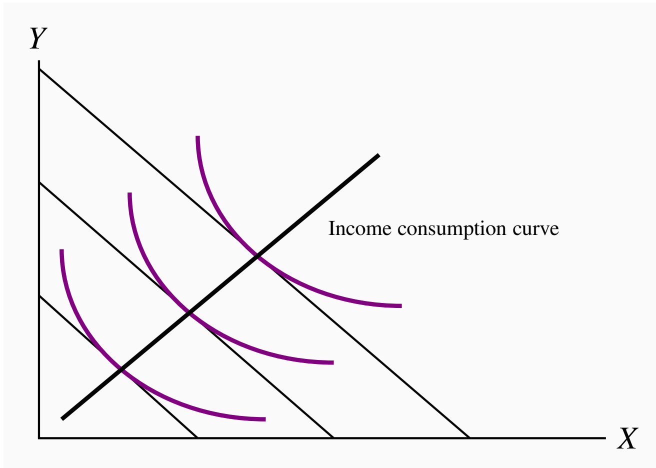


Exercise 6.8

They are not strictly convex to the origin, and so they do not display a diminishing marginal rate of substitution.

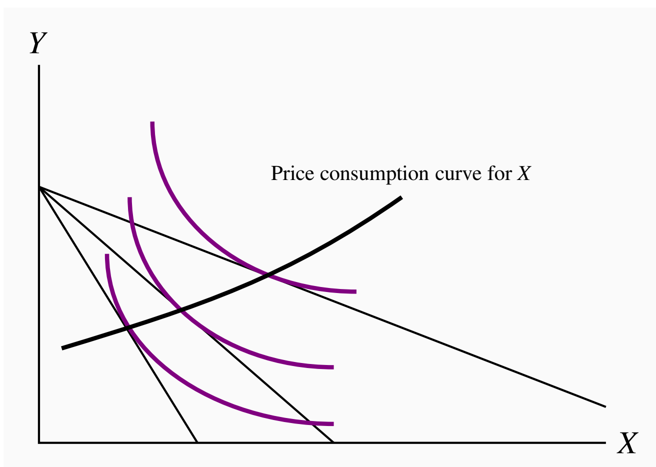
Exercise 6.9

See figure below.



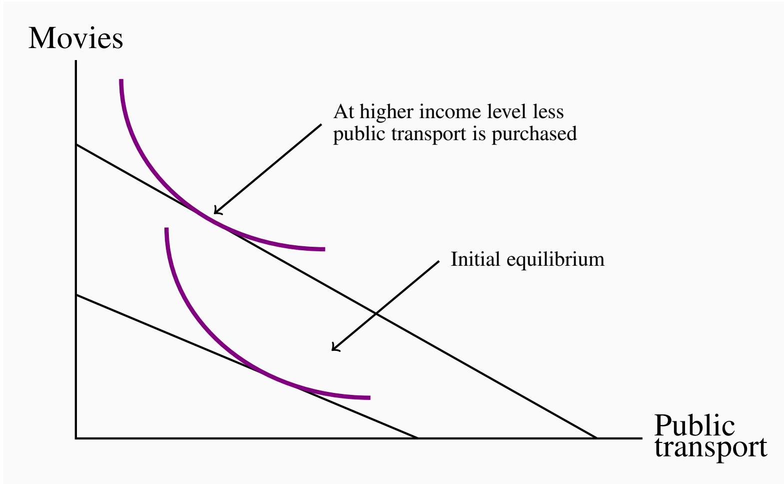
Exercise 6.10

See the figure below. Part (b) will see the rotation point stay at the X intercept.



Exercise 6.11

With movies on the Y axis and public transport on the X, the higher income equilibrium will lie to the north-west of the lower income equilibrium.



Chapter 7 Solutions

Exercise 7.1

(a) Accounting profit is \$15,000 (= $\$55,000 - \$40,000$).

(b) Economic costs include opportunity cost of the investment (\$7,000) plus the additional earnings of \$10,000 (= $\$50,000 - \$40,000$). These additional costs of \$17,000 mean that economic profits are negative.

(c) No.

(d) 5%

Exercise 7.2

If there is no medical exam then it is probable that less healthy individuals will avail of it. Knowing this, the firm should choose its benefit/payout structure to reflect a high-cost clientele. It will have lower payouts and/or higher premiums. Therefore a healthy individual would likely not obtain favourable insurance terms.

Exercise 7.3

(a) Spread.

(b) Pooled.

(c) Spread.

(d) Spread and pooling.

Exercise 7.4

(a) \$700 ($= \$350,000/500$)

(b) She is averse to risk.

Exercise 7.5

(a) See the diagram below.

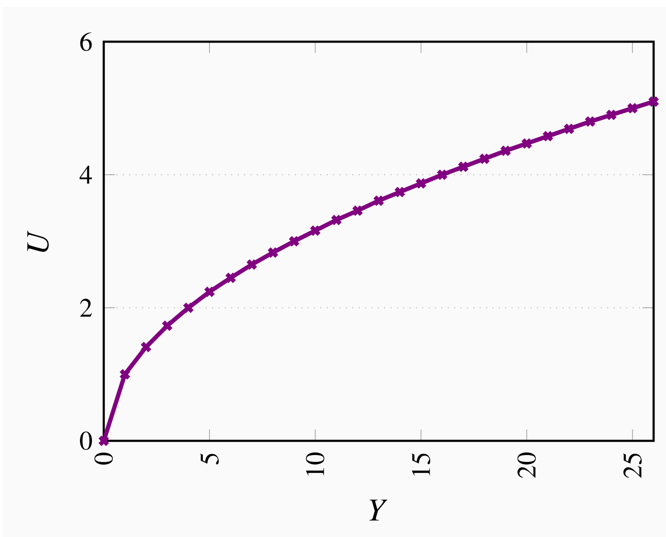
(b) See the diagram below.

(c) 2, 3, 4.

(d) 3 units ($=$ average of $\sqrt{4}+\sqrt{16}$).

(e) ($= \sqrt{10}$)

(f) This is because of diminishing marginal utility: the loss in utility in going from \$10 to \$4 exceeds the gain in utility from \$10 to \$16.



Exercise 7.6

(a) See the diagram below.

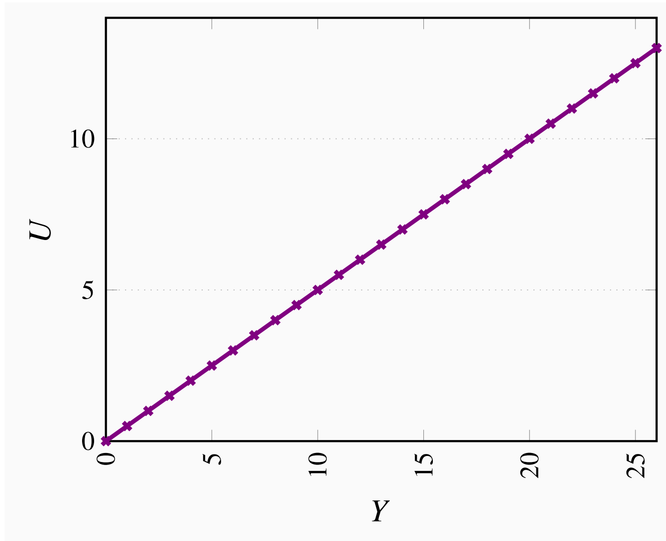
(b) See the diagram below.

(c) 2, 4.5, 8.

(d) 5 ($=$ average of $2+8$).

(e) 5

(f) In this case the MU of money is constant. Each increment in income yields the same additional utility. Hence the utility loss of -\$6 equals the utility gain of +\$6.



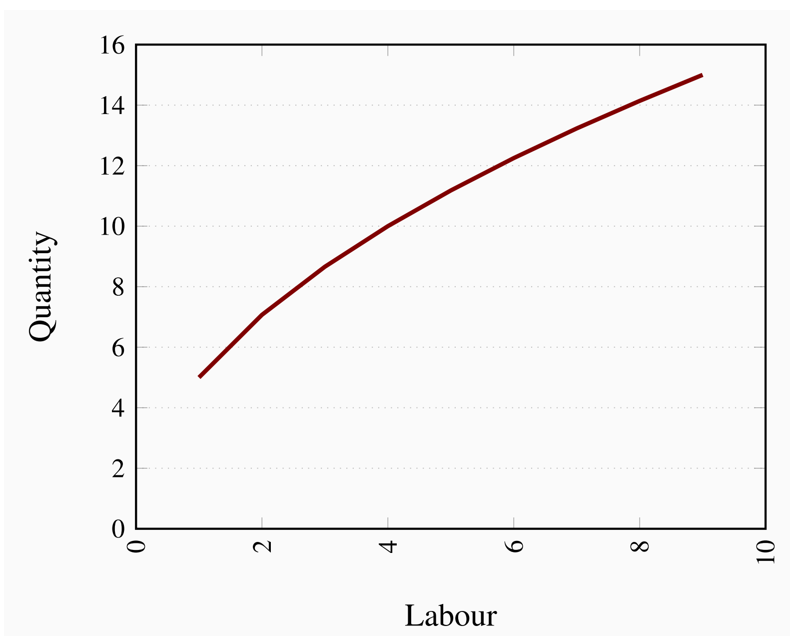
Chapter 8 Solutions

Exercise 8.1

(a) The MP is the difference in output at different labour levels: 5, 2.07, 1.59, 1.34, 1.18,. . .

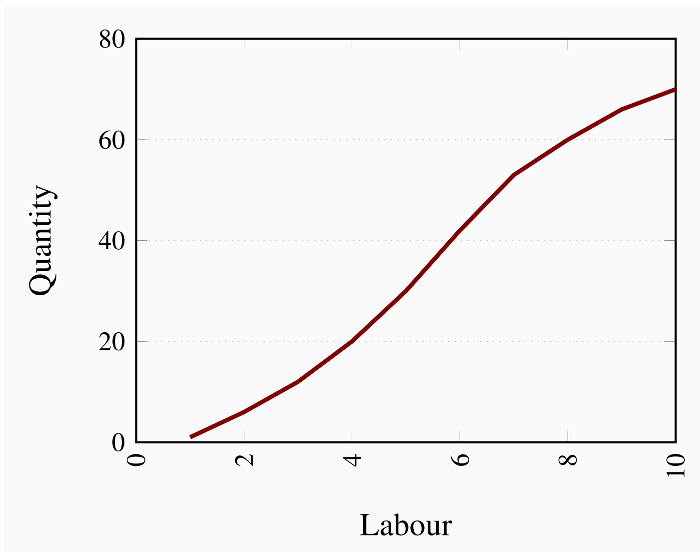
(b) See the figure below.

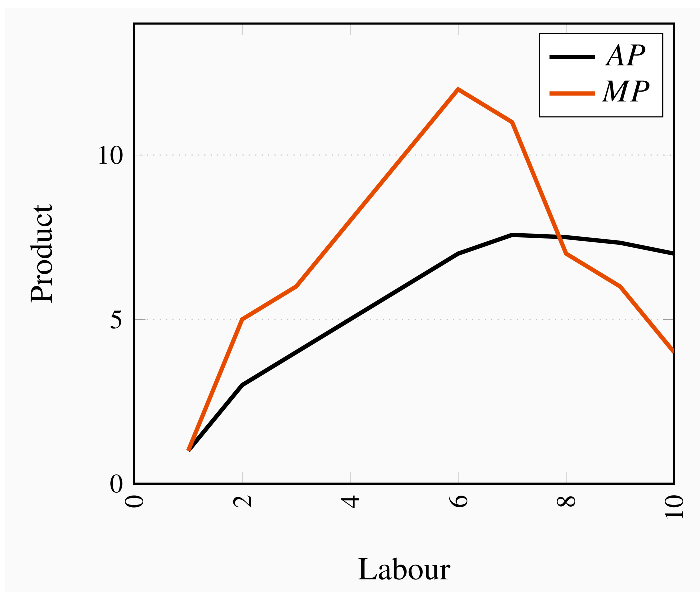
(c) Note that total output increases at a diminishing rate – the MP is declining.



Exercise 8.2

For each level of labour used, its *AP* is: 1.0, 3.0, 4.0, 5.0, 6.0, 7.0, 7.57, 7.5, 7.33, 7.0. and the *MP* is: 1, 5, 6, 8, 10, 12, 11, 7, 6, 4. The *AP* and *MP* are graphed below. If the *MP* cuts the *AP* at the latter's maximum, your graph is likely correct.





Exercise 8.3

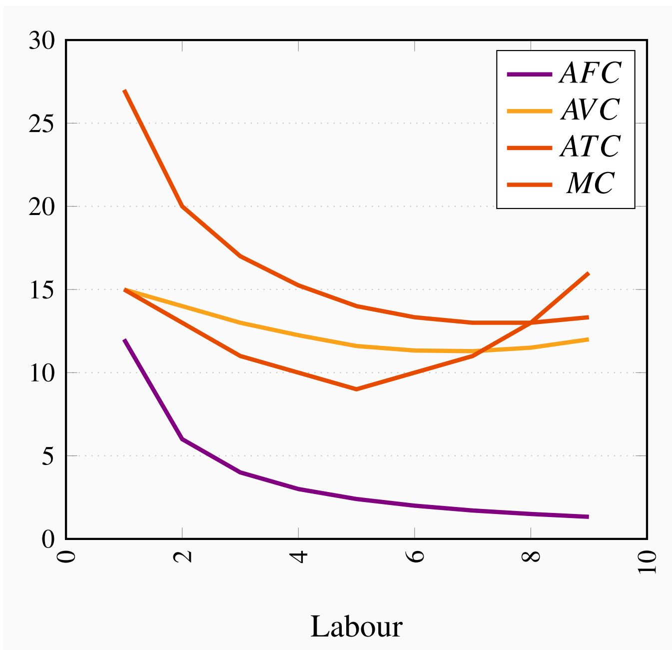
(a) Fixed cost is $12.

(b) See below.

(c) See below.

Q	TC	AFC	AVC	ATC	MC
0	12				
1	27	12.00	15.00	27.00	15
2	40	6.00	14.00	20.00	13
3	51	4.00	13.00	17.00	11
4	61	3.00	12.25	15.25	10
5	70	2.40	11.60	14.00	9
6	80	2.00	11.33	13.33	10
7	91	1.71	11.29	13.00	11
8	104	1.50	11.50	13.00	13
9	120	1.33	12.00	13.33	16

(d) See below.

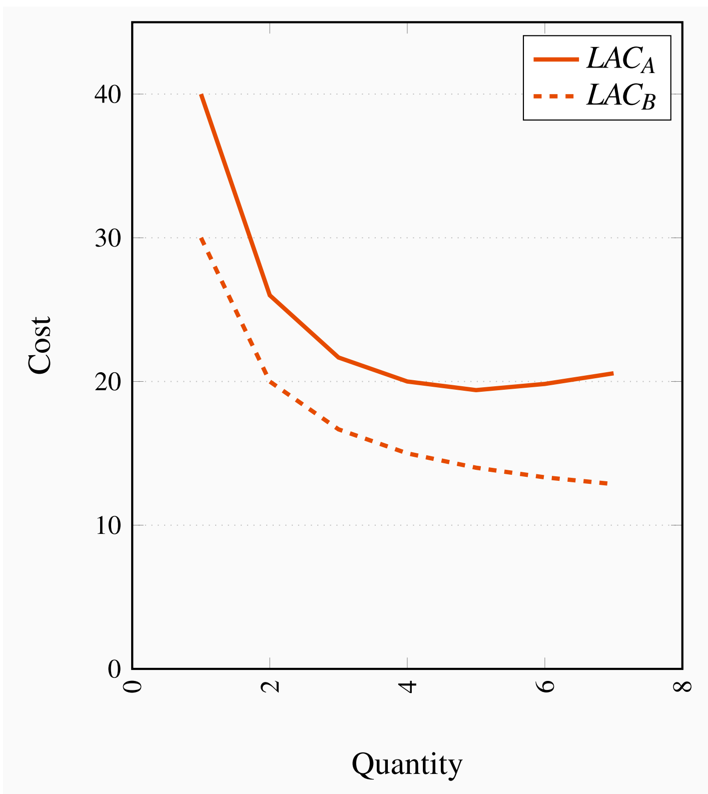


Exercise 8.4

(a) The costs are given in the table below.

(b) Firm A experiences decreasing returns to scale at high outputs, whereas B does not.

Q	TC_A	LAC_A	TC_B	LAC_B
1	40	40.00	30.00	30.00
2	52	26.00	40.00	20.00
3	65	21.67	50.00	16.67
4	80	20.00	60.00	15.00
5	97	19.40	70.00	14.00
6	119	19.83	80.00	13.33
7	144	20.57	90.00	12.86



Exercise 8.5

MC curve data are given in the table below. Firm B has constant marginal costs in the LR; hence never encounters decreasing returns to scale. Firm A's LR MC intersects its LR ATC at an output between 5 and 6 units, where the ATC is at a minimum. Firm B's MC lies everywhere below its ATC.

Output	1	2	3	4	5	6	7
***MC* Firm A**	40	12	13	15	17	22	25
***MC* Firm B**	30	10	10	10	10	10	10

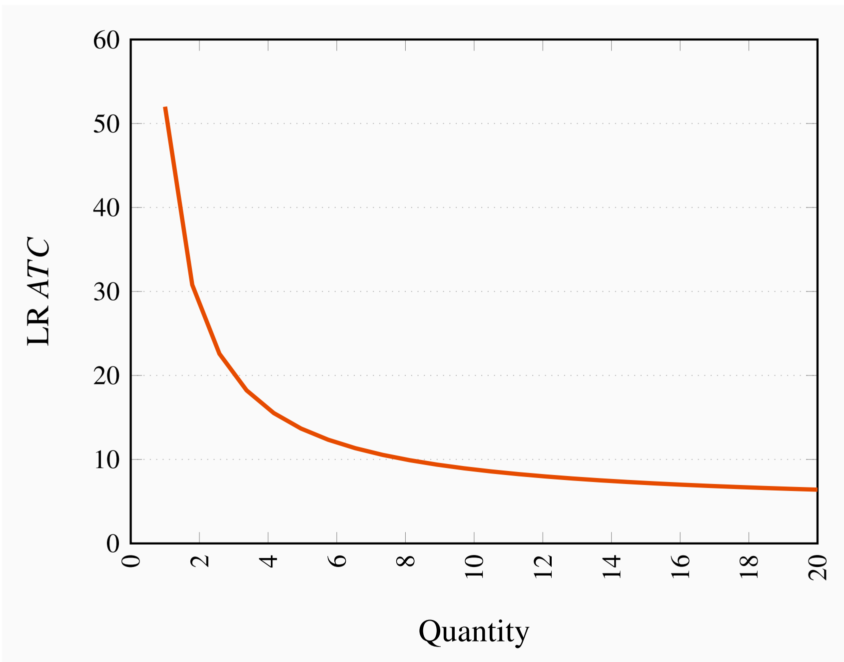
Exercise 8.6

(a) See graphic below.

(b) Decreasing returns to scale.

(c) As q becomes infinitely large the second term tends to zero, hence ATC tends to \$4.

(d) $LMC = 4$.



Chapter 9 Solutions

Exercise 9.1

(a) The MC is \$32.

(b) Her break-even level of output is 25 units.

(c) No, because she can cover her variable costs. $TVC = \$320$; $TR = \$360$.

Exercise 9.2

For total revenue to equal total cost it must be the case that $130 \times Q = 200,000 + 80 \times Q$. Therefore $Q = 4,000$.

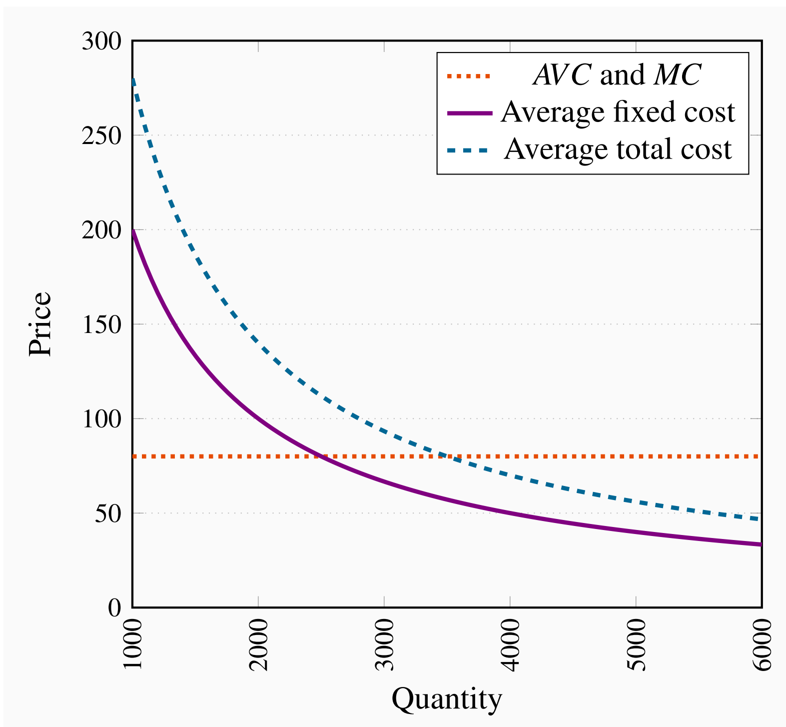
Exercise 9.3

(a) The MC is horizontal at \$80.

(b) See diagram below.

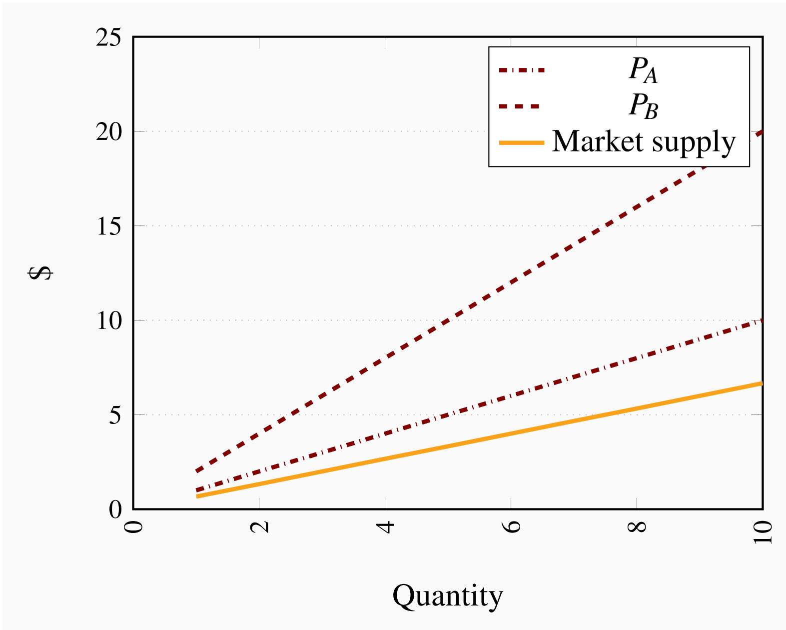
(c) See diagram below.

(d) The MC would have to increase at some point.



Exercise 9.4

The market supply curve goes through the origin with a slope of 2/3. This follows from the fact that we can write the supply curves as $q_A = P$ and $q_B = 0.5P$. Hence $Q = q_A + q_B = 1.5P$; or $P = (2/3)Q$.



Exercise 9.5

Since the costs per unit are declining with output, they are producing on the downward-sloping segment of the $LATC$. To see this we need just calculate ATC at each output.

Exercise 9.6

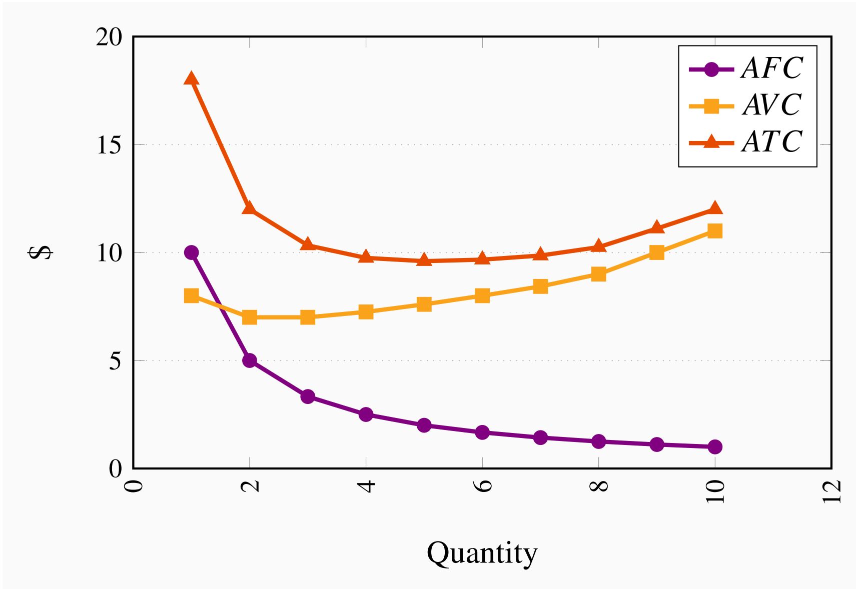
(a) See the table.

(b) See the figure below.

(c) $Q = 7$. At this output $MC = MR$.

(d) Price is fixed.

Q	0	1	2	3	4	5	6	7	8	9	10
TC	10	18	24	31	39	48	58	69	82	100	120
TR	0	11	22	33	44	55	66	77	88	99	110
Profit	−10.00	−7.00	−2.00	2.00	5.00	7.00	8.00	8.00	6.00	−1.00	−10.00
MR		11.00	11.00	11.00	11.00	11.00	11.00	11.00	11.00	11.00	11.00
MC		8.00	6.00	7.00	8.00	9.00	10.00	11.00	13.00	18.00	20.00
AFC		10.00	5.00	3.33	2.50	2.00	1.67	1.43	1.25	1.11	1.00
AVC		8.00	7.00	7.00	7.25	7.60	8.00	8.43	9.00	10.00	11.00
ATC		18.00	12.00	10.33	9.75	9.60	9.67	9.86	10.25	11.11	12.00



Exercise 9.7

(a) The equilibrium price is \$40 and the equilibrium quantity is 6,000. The price intercept for the demand equation is 50, the quantity intercept 30,000. The price intercept for the supply equation is 10 and the quantity intercept -2,000.

(b) With a perfectly competitive structure, this new firm cannot influence the price. Therefore it maximizes profit by setting $P = MC$. That is $40 = 10 + 0.5q$. Solving this equation yields a quantity value $q = 60$.

Exercise 9.8

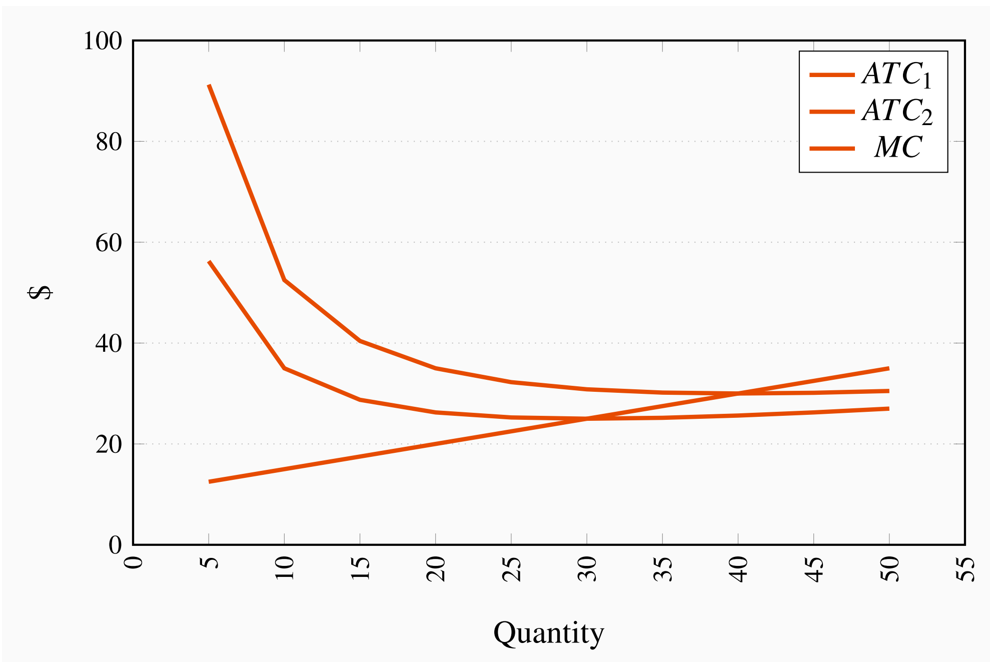
(a) See the table below.

(b) See the diagram below.

(c) The break-even prices correspond to the minimum of each ATC curve: \$30 and \$25.

(d) The price in the market will be forced down to the level at which the most efficient producers can supply the market. Consequently the producer with the higher fixed cost will either have to adopt the technology of the lower-cost producer or exit the industry.

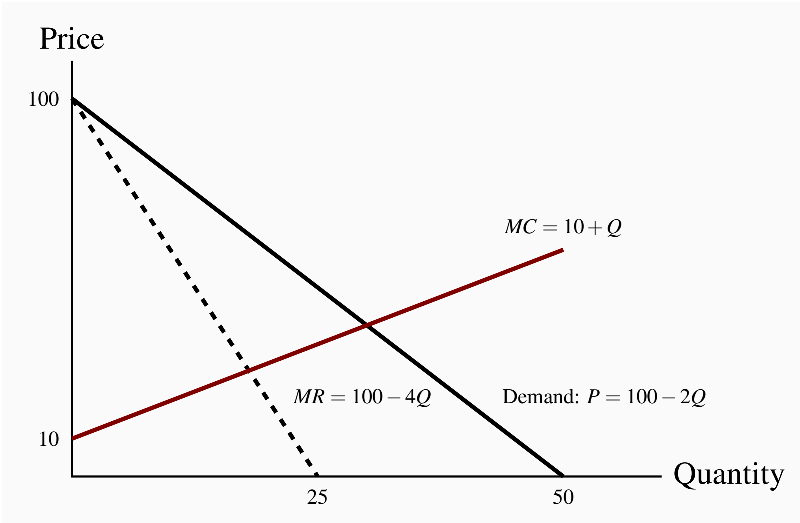
Q	ATC_1	ATC_2	MC
5	91.25	56.25	12.5
10	52.50	35.00	15
15	40.42	28.75	17.5
20	35.00	26.25	20
25	32.25	25.25	22.5
30	30.83	25.00	25
35	30.18	25.18	27.5
40	30.00	25.63	30
45	30.14	26.25	32.5
50	30.50	27.00	35



Chapter 10 Solutions

Exercise 10.1

This is a standard diagram for the monopolist. See the figure below. Equating $MC = MR$ yields $Q = 18$, $P = \$64$.

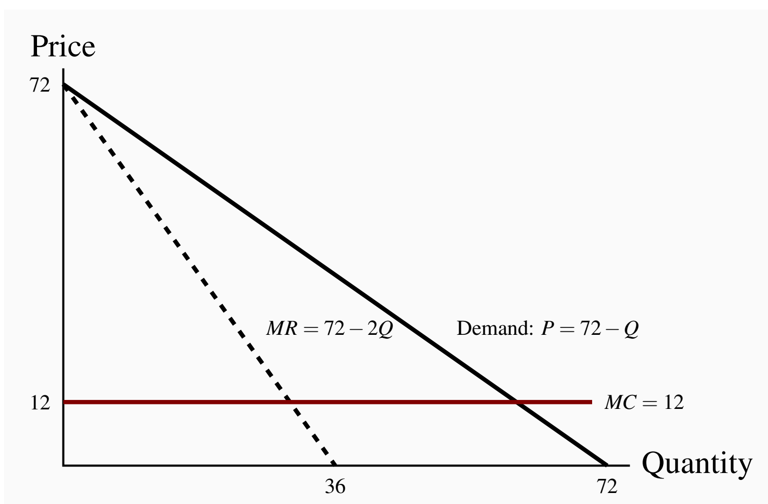


Exercise 10.2

(a) See below.

(b) Total revenue is a maximum where MR becomes zero. This is at $P = \$36$ and $Q = 36$.

(c) The quantity $Q = 36$ sells at a price $P = \$36$, from the demand curve. Hence total revenue is $\$36 \times 36 = \$1,296$.



Exercise 10.3

(a) Where demand equals MC, we obtain $72 - Q = 12$. Therefore $Q = 60$.

(b) From the demand curve, if $Q = 60$ then $P = \$12$.

(c) The efficiency gain in going from a profit maximizing monopoly ($Q = 30$) to perfect competition ($Q = 60$) is given by the area under the demand curve and above the MC curve between these output levels. This is $1/2 \times 30 \times 30 = \$450$.

Exercise 10.4

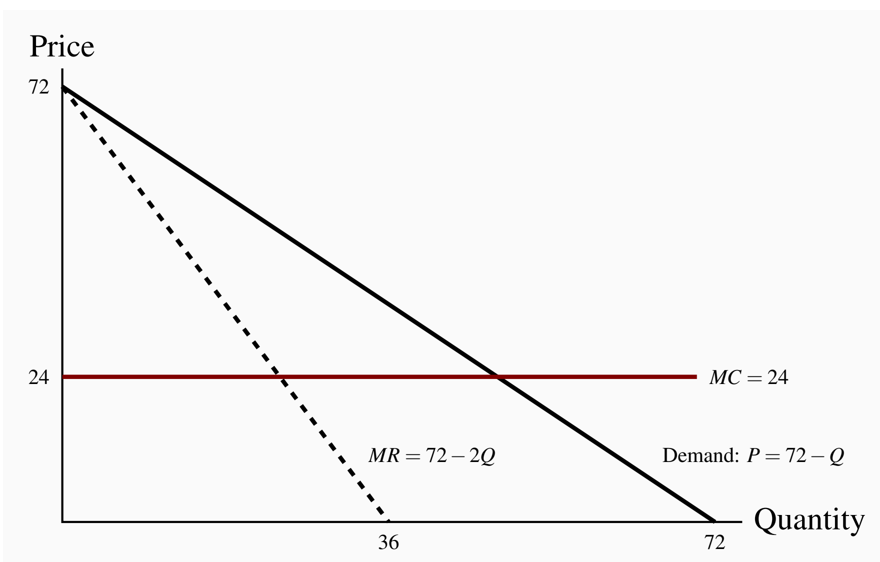
The buyers' reservation prices are given in row P. The cost of producing each unit is given in row MC. The profit on the first unit is therefore $\$(14-2) = \$12$; on the second unit is $\$(12-3) = \$9$, etc. On the fifth unit the additional profit is zero. Therefore, four units should be produced and sold. Total profit is $\$12+\$9+\$6+\$3 = \$30$.

Exercise 10.5

(a) See figure below. The profit maximizing outcome is $Q = 24$ and $P = \$48$ – obtained from $MC = MR$.

(b) With perfect price discrimination the monopolist's revenue is the area under the demand curve. He should continue to produce and sell as long as the demand price is greater than MC. Where the demand price equals MC profit is maximized. This occurs at $P = \$24$, $Q = 48$.

(c) Profit is $TR - TC$ at $Q = 24$. This is \$576. In (b) the profit is the area under the demand curve up to the output $Q = 48$ minus the area under the MC curve up to this same output. This is \$1152.



Exercise 10.6

(a) Profit maximizing output is where $MC = MR$ in each market. The MRs are $MR_A = 20 - (1/2)Q_A$, $MR_B = 14 - (1/2)Q_B$. Equating each of these in turn to $MC = 4$ yields $Q_A = 32$

and $Q_B = 20$. These outputs can be sold at a price obtained from their demand curves: $P_A = \$12$ and $P_B = \$9$. This is illustrated in the table below.

(b) Total profit is the sum of profit in each market: $(P_A \times Q_A - TC_A) + (P_B \times Q_B - TC_B) = \$256 + \$100 = \$356$.

Q	P_A	P_B	MC	MR_A	MR_B
1	19.75	13.75	4	19.5	13.5
4	19	13	4	18	12
8	18	12	4	16	10
12	17	11	4	14	8
16	16	10	4	12	6
20	15	9	4	10	4
24	14	8	4	8	2
28	13	7	4	6	0
32	12	6	4	4	-2
36	11	5	4	2	-4
40	10	4	4	0	-6
44	9	3	4	-2	-8
48	8	2	4	-4	-10
52	7	1	4	-6	-12
56	6	0	4	-8	-14
60	5	-1	4	-10	-16
64	4	-2	4	-12	-18
68	3	-3	4	-14	-20
72	2	-4	4	-16	-22
76	1	-5	4	-18	-24

Exercise 10.7

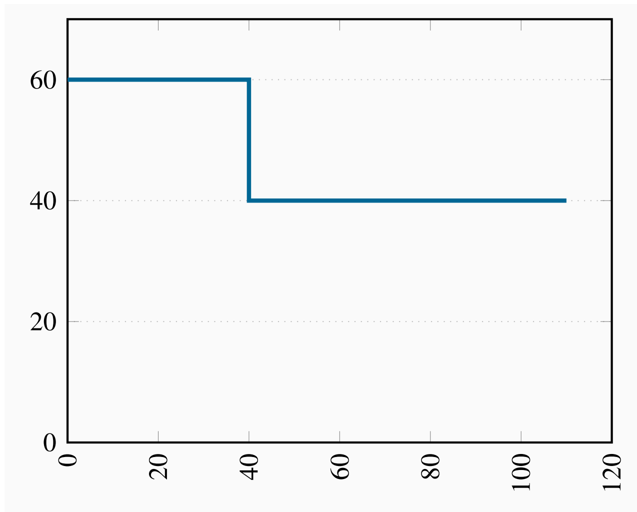
(a) The diagram here is equivalent to the one in Figure 10.11 in the text. The first segment, up to an output of 40 units, has a price of \$60; the second, from an output of 40 to 110, has a price of \$40.

(b) The *MC* curve runs along the horizontal axis – after the fixed cost is incurred, the *MC* is zero. The demand curve is the *MR* curve here, composed of the two horizontal segments.

(c) A price of \$60 can be charged to 40 buyers, and a price of \$40 charged to 70 buyers. Hence

$TR = \$5,200$. Since $TC = \$3,500$, profit is \$1,700.

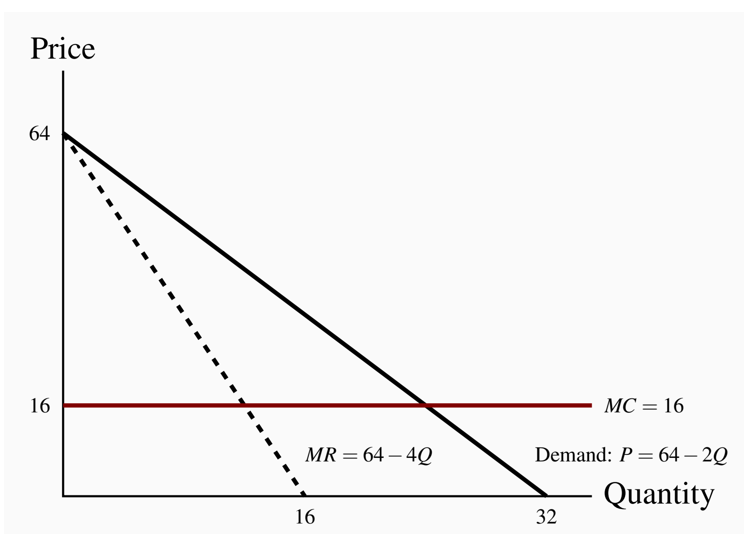
(d) Yes; 110 buyers at \$40 each yields a $TR = \$4,400$. Subtract the TC to yield a profit of \$900.



Exercise 10.8

(a) Setting $MC = MR$ yields $Q = 12$ and from the demand curve, $P = \$40$. See the figure below.

(b) Where MC equals demand the output is $Q = 24$. In moving from the output level $Q = 24$ to $Q = 12$, the DWL is the area bounded by the demand curve and the MC between these output levels: $1/2 \times 12 \times 24 = \$144$.



Chapter 11 Solutions

Exercise 11.1

The three-firm ratios are 0.14, 0.49, 0.80, 0.94. The four-firm ratios are 0.15, 0.54, 0.82, 0.95.

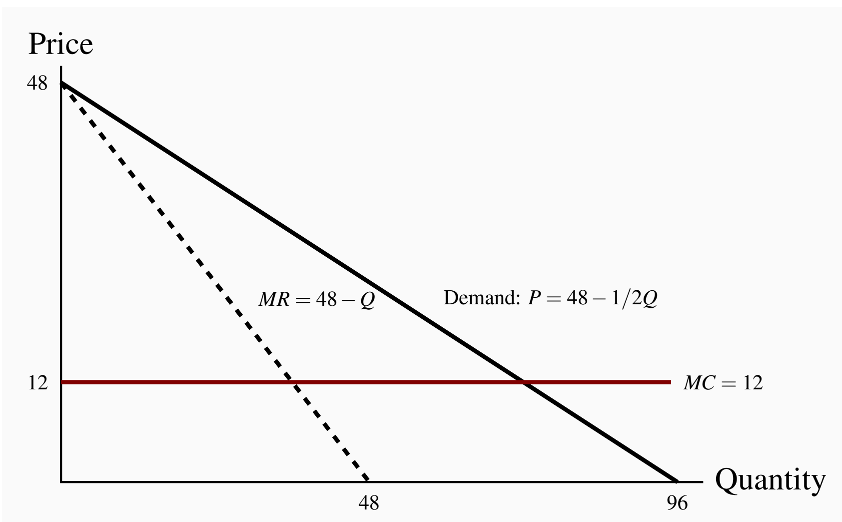
Exercise 11.2

(a) See graph below.

(b) Equating MC to MR yields $Q = 36$ and therefore, from the demand curve, $P = \$30$ when $Q = 30$.

(c) TR is \$1,080, total cost is \$432 and therefore profit is \$648.

(d) Profits plus freedom of entry will see new firms take some of this firm's market share, and therefore reduce profit.



Exercise 11.3

(a) The diagram here is similar to the one above.

(b) Acting as a monopolist they would set $MR = MC$, hence $Q = 90$, $P = \$70$.

(c) Combined profit is $\$90 \times (70 - 40) = \$2,700$. Individual profit is half of this amount.

Exercise 11.4

(a) Yes. If A confesses then B's best strategy is also to confess. If A denies, B's best strategy is also to confess. Hence, either way B's best choice is to confess – this is a dominant strategy.

The same reasoning applies to A.

(b) The Nash Equilibrium is that they both confess.

(c) Yes.

(d) If the crooks could communicate with each other they could cooperate and agree to deny. This would be better for each.

Exercise 11.5

(a) Each firm has a 'high output' dominant strategy, since their profit is greater here regardless of the output chosen by the other firm.

(b) From (a) it follows that high/high is the Nash Equilibrium.

(c) Since low/low yields more profit for each firm, a cartel is an attractive possibility. But it may not be sustainable, given that each player has the incentive to renege on the cartel agreement.

Exercise 11.6

(a) While Ronnie can threaten to lower its price if Flash enters the market it would not be profitable for Ronnie to do that because a higher price, even with Flash in the market, yields a superior profit to Ronnie. Hence Flash should enter.

(b) The issue here is that the threat to lower price is not credible.

Exercise 11.7

(a) Equating price to MC yields $Q = 11,200$.

(b) Equating MR to MC yields $Q = 5,600$.

(c) Using the formula $Q = n/(n+1) \times$ (perfectly competitive output) yields market $Q = 2/3 \times 11,200 = 7,466.67$.

Exercise 11.8

(a) Profit under perfect competition is zero (only normal profit). Under monopoly the price charged is \$1,800. Cost per unit is \$400, and quantity produced is 5,600. Hence profit $= 5,600 \times (1,800 - 400) =$ \$7.84m. Since the output in the duopoly market is 2/3 times the perfectly competitive output, then $Q = 7,466.67$. The price is thus $P = 3,200 - (1/4) \times 7,466.67 = \$1,333.33$. Profit per unit is thus \$933.33, and total profit is \$6.97m.

(b) Since the unit costs are constant we could have any number of firms producing in this market.

Chapter 12 Solutions

Exercise 12.1

At $L = 4$ the VMP of labour is \$400, which is also the wage rate. Therefore this is the profit maximizing output level (on the assumption that fixed costs are less than profits). The table below contains the calculations.

Labour	Output	MP Labour	VMP Labour	TR	TC	Profit
0	0					
1	20	20	400	400	400	0
2	50	30	600	1000	800	200
3	75	25	500	1500	1200	300
4	95	20	400	1900	1600	300
5	110	15	300	2200	2000	200
6	120	10	200	2400	2400	0

Exercise 12.2

Here you must calculate the additional cost of each employee. The first costs \$250; the second \$350 (\$300 plus an additional \$50 to the first employee); the third \$450; the fourth \$550, etc. The additional revenue from each employee is the VMP of labour. As long as this exceeds the MC of hiring another employee then that employee should be hired. The answer is thus $L = 3$.

Labour	Output	MP Labour	VMP Labour	Marginal Wage	MC Labour
0	0				
1	20	20	400	250	250
2	50	30	600	300	350
3	75	25	500	350	450
4	95	20	400	400	550
5	110	15	300	450	650

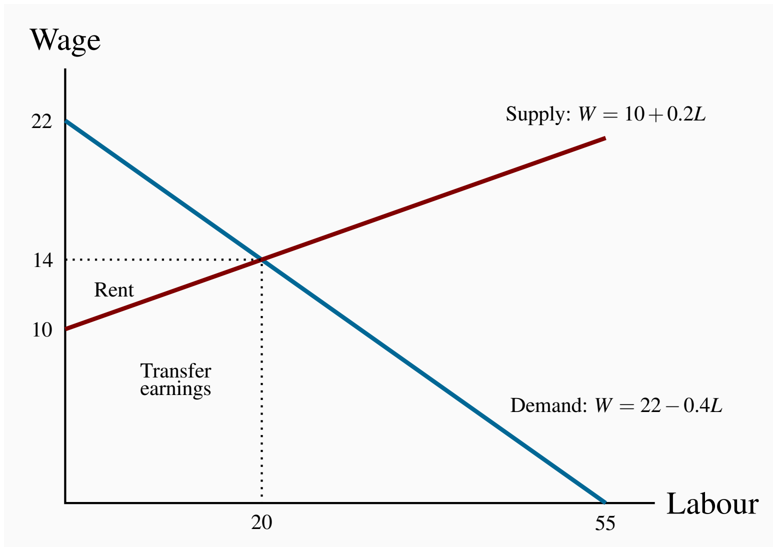
Exercise 12.3

(a) In a spreadsheet, for each value of L, we can compute a supply price and a demand price. Those prices are equal at $L = 20$, $W = \$14$.

(b) See the diagram below.

(c) See the diagram below.

(d) Total wage bill is \$280, of which transfer earnings account for \$240 and rent \$40.



Exercise 12.4

The present values of the three streams are: \$16,333.3; \$943.4; \$8,545.5. Therefore only the first project should be adopted because it alone generates revenue in excess of costs.

Exercise 12.5

Only the first plot generates sufficient revenue to yield a profit.

Exercise 12.6

(a) You would hold it for 5 years because the wine is appreciating by more than the cost of borrowing for each of the first five years. The sixth year the wine grows in value by the same as the borrowing cost.

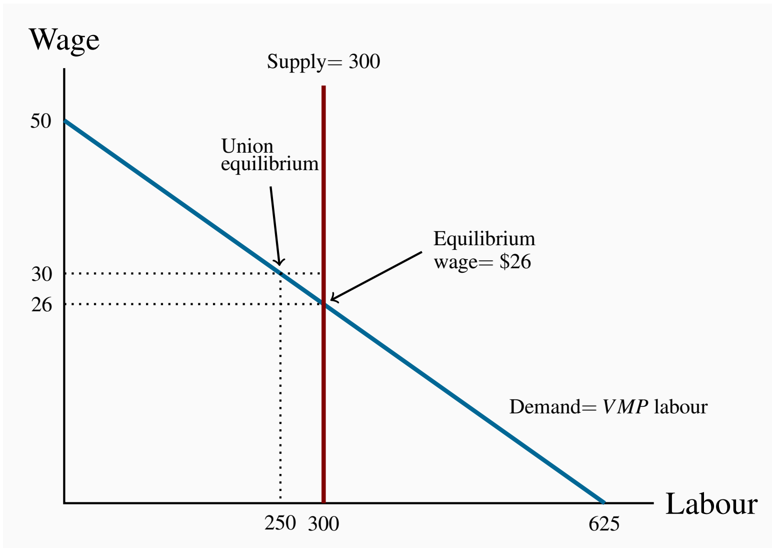
(b) In this case the carrying has increased to 7% per year. So it would be profitable to hold the wine for three years – until the growth in the value of the wine equals the carrying cost.

Exercise 12.7

(a) The demand curve has a regular downward-sloping form, while the supply curve is vertical at $L = 300$. See figure below.

(b) At $L = 300$ the demand curve indicates that the wage is \$26.

(c) At $W = \$30$, the corresponding demand is $L = 250$.



CHAPTER 13 SOLUTIONS

Exercise 13.1

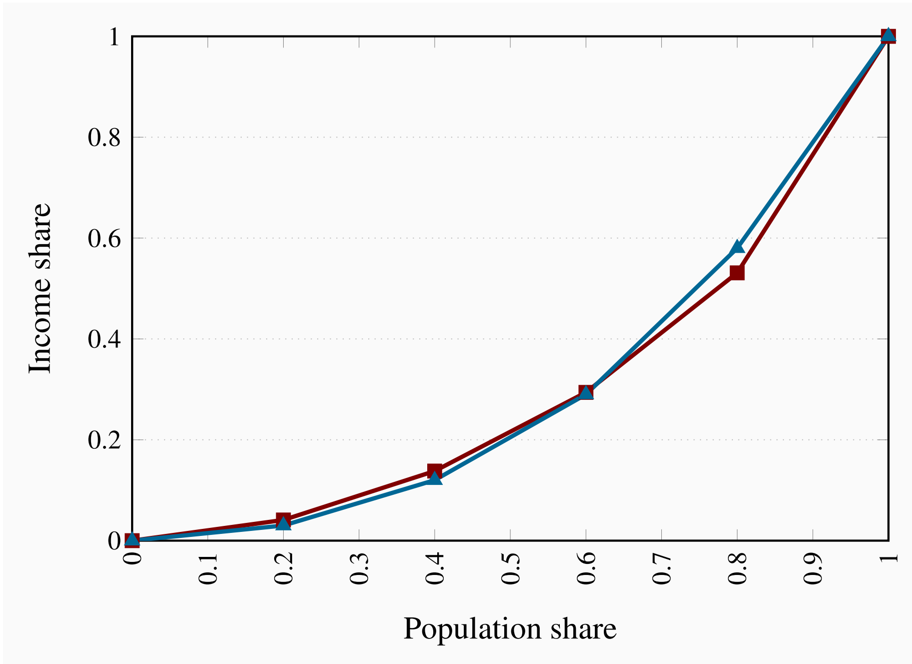
(a) The present value of going to university is higher at an interest rate of 10%. If you discount the first stream of values you will obtain $-20,000$, 36,360 and 41,320 yielding a net present value of 57,680. With 20,000 dollars each period in contrast, the net present value is 54,710 dollars.

(b) By performing the same set of calculations using the 2% discount rate, you will find that university is still preferred.

(c) At an interest rate above 15% the 'no university' option will yield a higher net present value. Try discounting the two income streams using a rate of 16% and you will see.

Exercise 13.2

These two distributions have intersecting Lorenz curves, so it is difficult to say which is more unequal without further analysis.

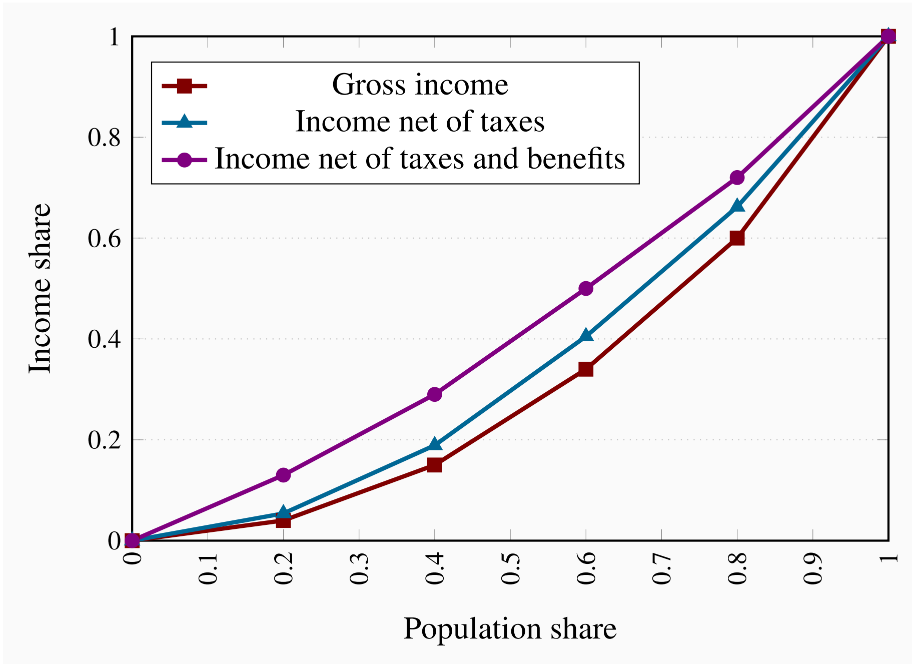


Exercise 13.3

(a) The coordinates on the vertical axis measured in percentages are: 4, 15, 34, 60, 100. See the figure below for the graphic.

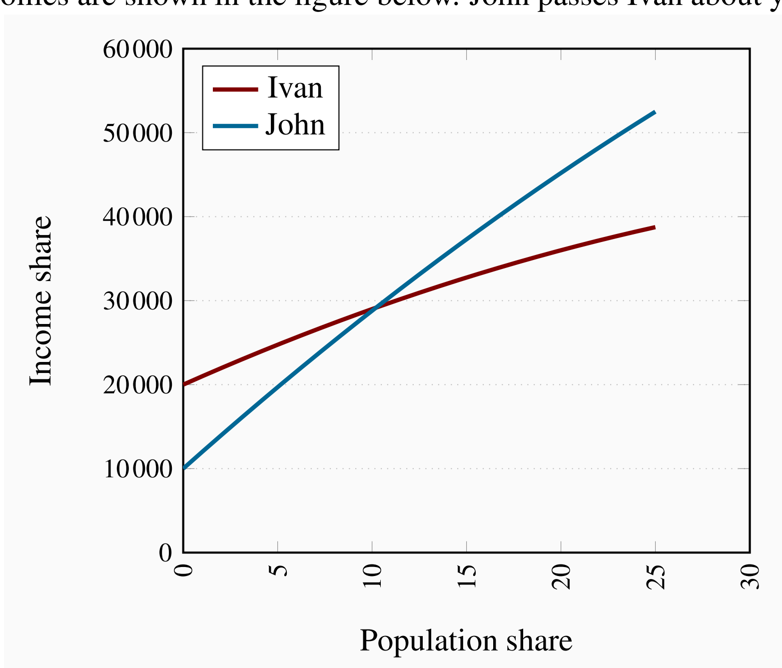
(b) The new coordinates are: 5.4, 18.9, 40.5, 66.2, 100.

(c) The coordinates for post government income are: 13, 29, 50, 72, 100. The three Lorenz curves are plotted below.



Exercise 13.4

The profiles are shown in the figure below. John passes Ivan about year ten.



Exercise 13.5

(a) Solving the demand and supply involves equating $Q = 40 - 40f$ to $f = 0.5$. Thus the equilibrium premium is 20, which is interpreted in percentage terms. See figure below.

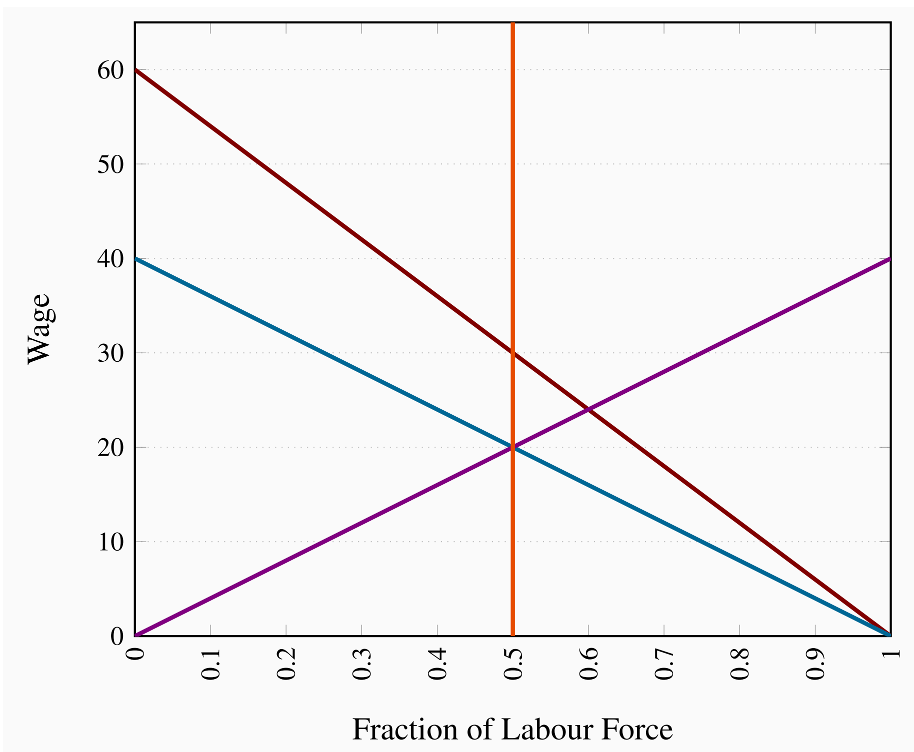
(b) If demand shifts upwards to $W = 60 - 60f$, the new equilibrium is 30 percent, as illustrated in the figure below again.

Exercise 13.6

(a) See diagram below.

(b) See diagram below.

(c) In the long run the relative supply is $W = 40f$, and equating this with demand yields a 24 percent premium rather than a 30 percent premium and $f = 0.6$.



Chapter 14 Solutions

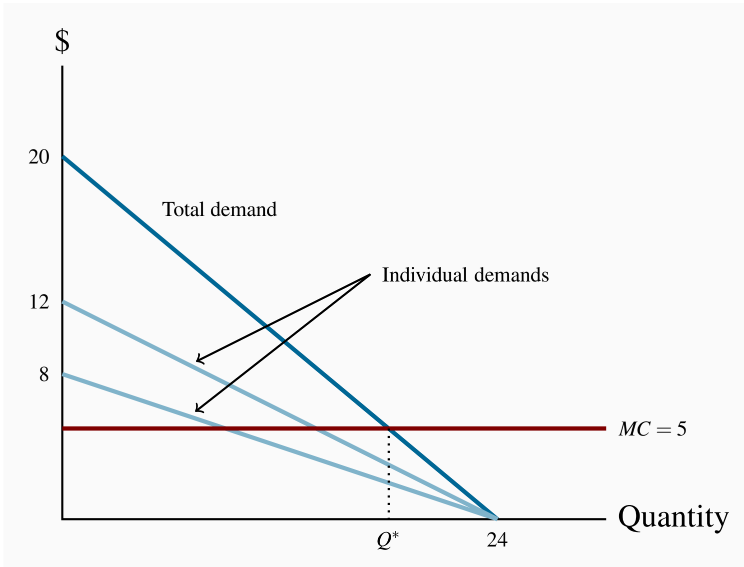
Exercise 14.1

(a) See figure below.

(b) The total demand for the public good has a vertical intercept of 20 and a horizontal intercept of 24. The form of the equation is therefore $P = 20 - (5/6)Q$.

(c) Equate the MC of \$5 to the total demand curve to obtain $Q = 18$. This is the 'optimal' output – where the cost of the last unit produced equals the value placed on it by both individuals. At this quantity the individual valuations (the price that each is willing to pay) are obtained from the individual demand curves. Substituting $Q = 18$ into each yields \$3 and \$2.

(d) The area below the total demand, above the quantity axis up to Q^*.



Exercise 14.2

(a) The new individual demand has a value of \$10 and a quantity intercept of 24. The new total demand has intercepts $\{\$30, 24\}$.

(b) The new Q^* is greater than with two individuals. Hence the total value of the public good rises.

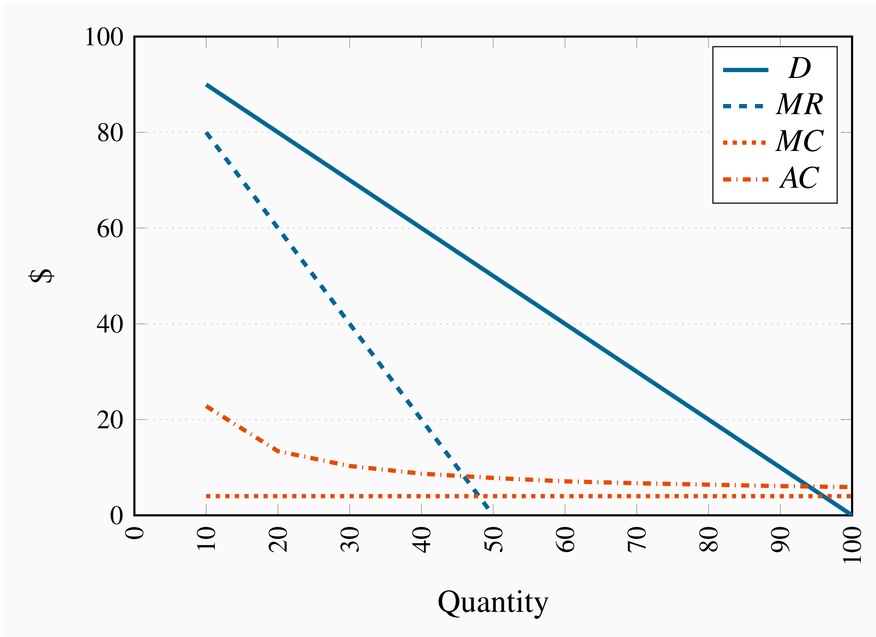
(c) See the diagram.

Exercise 14.3

(a) See the diagram below.

(b) The efficient output is where the $MC = P$, as given by the demand curve. He would maximize profit by producing where $MC = MR$.

(c) He would choose a price from the demand curve at an output where $MC = MR$.



Exercise 14.4

(a) Equating the ATC to the demand curve yields $100 - Q = 4 + 188/Q$. The solution is $Q = 94$.

(b) This is (slightly) less than the efficient output.

(c) Average cost is \$(4+188)/94=\$6; price is \$6.

(d) The DWL under this regulation equals the area beneath the demand curve and above the MC curve between quantity values 94 and 96. The unrestricted monopoly DWL is the same area between the quantity values where $MC = MR$ and 96.

Exercise 14.5

An efficient output is where $P = MC$, that is $Q = 96$. At this output he charges a price of \$4. Hence his loss per unit is the difference between price and ATC which is 188/96. Since he produces 96 units, then charging a price of just \$4 leads to a revenue shortfall of $96 \times (188/96) = \$188$. This amount would have to be spread as a charge over the number of buyers in the market as a fixed cost associated with purchasing. In essence each buyer would have to pay a certain entry fee just to purchase the good.

Chapter 15 Solutions

Exercise 15.1

(a) Northland has an absolute advantage in the production of both goods, as it has lower labour requirements for each.

(b) The opportunity cost of 1 bushel of wheat is 1/2 litre of wine in Northland and 3/4 litre of wine in Southland.

(c) Northland has a comparative advantage in wheat while Southland does in wine.

(d) By reducing wheat production by 1 bushel, Southland can produce an additional 3/4 litre of wine.

(e) Both countries can gain if Northland shifts production from wine to wheat and the countries trade wine for wheat at a rate between 1/2 litre of wine for 1 bushel of wheat and 3/4 litre of wine for one bushel of wheat.

(f) By reducing wine production by 1/2 litre, Northland can increase wheat production by 1 bushel, which, at Southland's opportunity cost, exchanges for 3/4 litre of wine, giving Northland a gain of 1/4 litre of wine.

Exercise 15.2

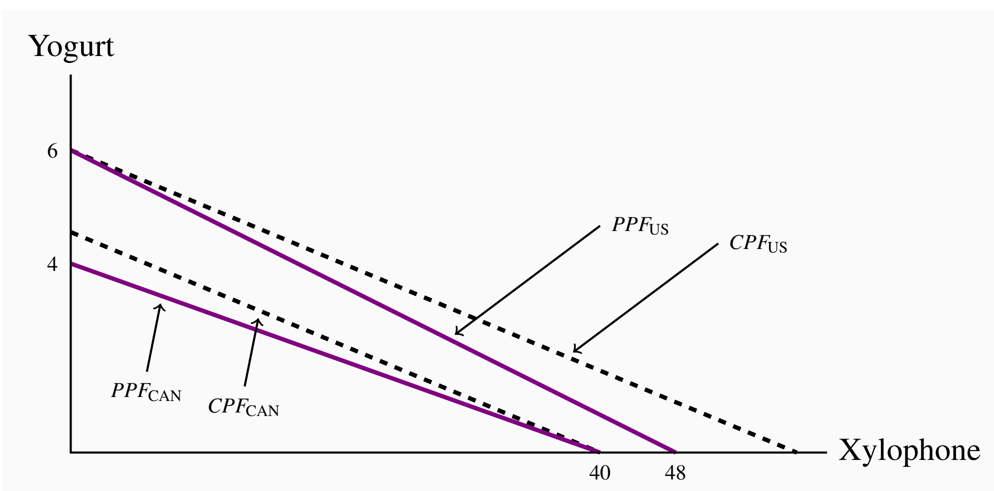
(a) The US has an absolute advantage in both goods.

(b) Canada has a comparative advantage in xylophones. The US has a comparative advantage in yogurt.

(c) See diagram below.

(d) See diagram below.

(e) See diagram below.



Exercise 15.3

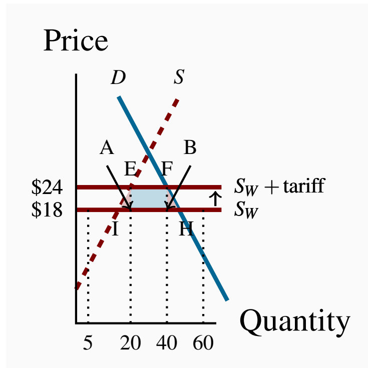
(a) The diagram shows that the amount traded is 60 units; of which domestic producers supply

5 and 55 are imported.

(b) In this case, the foreign supply curve SW shifts up from a price of \$18 to \$24. The amount traded is now 40 units, 20 of which are supplied domestically.

(c) Tariff revenue is EFHI= \$120.

(d) Equate the demand and supply curves to yield the above values.



Exercise 15.4

(a) The deadweight losses correspond to the two triangles, A and B, in the diagram, and amount to \$105.

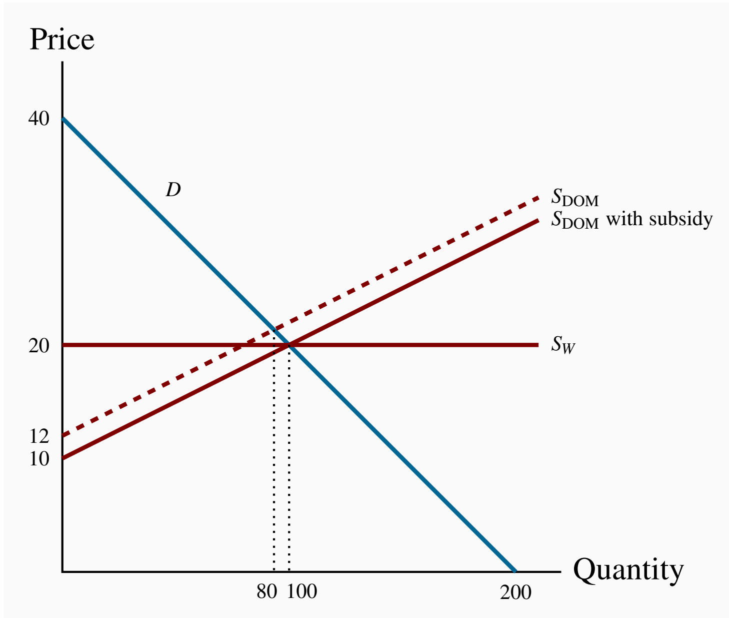
(b) The amount of additional profit for domestic producers is given by the area above the supply curve between prices \$18 and \$24.

Exercise 15.5

(a) See figure below. The total quantity of trade is 100 units, of which 80 are supplied domestically.

(b) The subsidy shifts the domestic supply curve down by \$2 at each quantity. This supply intersects the demand curve at $Q = 100$. Foreign producers are squeezed out of the market completely.

(c) Cost to the government is \$200 (100 units each with a \$2 subsidy.

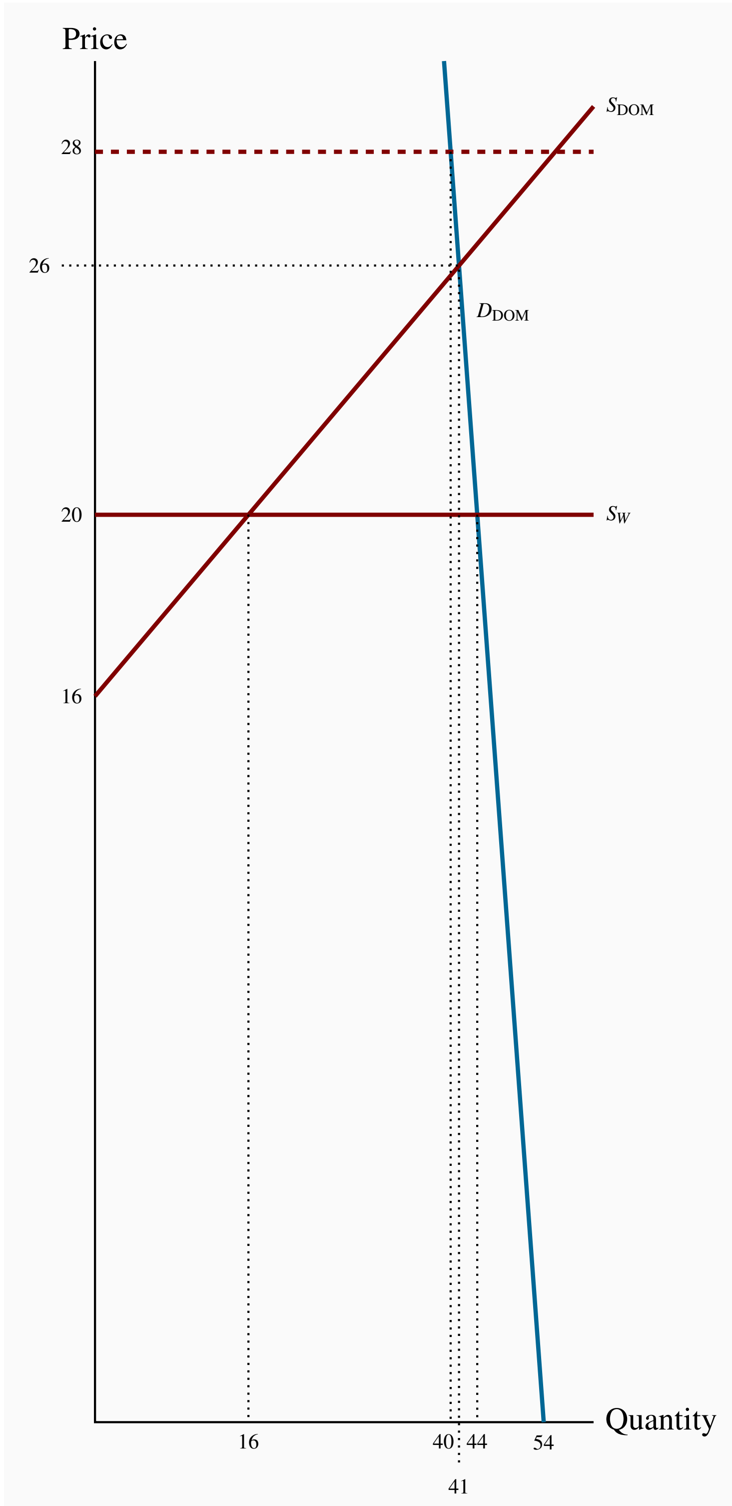


Exercise 15.6

(a) See diagram below. The US *PPF* now has an X intercept of 64 units.

(b) See diagram below.

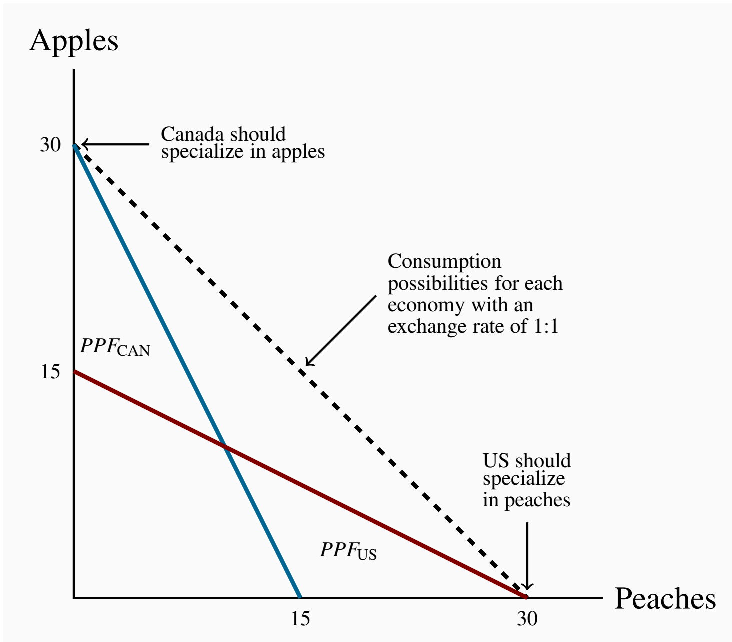
(c) Yes, the US should specialize in fish.



Exercise 15.7

The figure below illustrates parts (a) through (e).

For part (f): Since the total production before trade was 20 of each, and after specialization it is 30 of each, the gain is 10 of each good.



CHAPTER 16 SOLUTIONS

Exercise 16.1

(a) Rates of growth of real GDP:

$$\text{2012-13: } \left(\frac{1307}{1282} - 1\right) \times 100\% = 1.95\%\,,\ \text{2013-14: } -1.45\%$$

(b) Rates of inflation:

$$\text{2013: } \left(\frac{111.9}{109.1} - 1\right) \times 100\% = 2.7\%\,,\ \text{2014: } 24.1\%$$

(c) Rates of growth of labour force and employment:

2012-2013:

$$\text{Labour: } \left(\frac{17.857}{17.593} - 1\right) \times 100\% = 1.5\%\,,\ \text{Employ: } \left(\frac{16.696}{16.573} - 1\right) \times 100\% = 0.95\%$$

2013-2014:

Labour: 1.5%, Employ: 0.96%

(d) Unemployment rates 2010-2012:

$$2012: \left(\frac{17.593 - 16.537}{17.593}\right) \times 100\% = 6.0\%$$

$$2013: 6.5\%$$

$$2014: 7.0\%$$

Unemployment increased in 2011 and 2012 because the growth in employment was less than the growth in the labour force.

Exercise 16.2

(a) The participation and employment rates in 2014 were:

$$\left(\frac{18.125}{27.885}\right) \times 100\% = 65\% \text{ and } \left(\frac{16.856}{27.885}\right) \times 100\% = 60.4\%$$

(b) The labour force would decline to $(0.645 \times 27.885) = 17.986$ without any change in employment. As a result the unemployment rate in 2012 would fall to 6.3% but the employment rate would be unchanged. The fall in the participation rate lowers the size of the labour force. The population and employment, and the employment rate are unchanged.

Exercise 16.3

Value added is the difference between the market value of final output and the costs of intermediate inputs to production. In this case the market value of final output is \$1,000, the cost of inputs is \$625, i.e. $(\$350 + \$125 + \$150)$, and value added is \$375 $(\$1,000 - \$625)$. If brewers wholesale some of their output it is an intermediate input to the service provided by pubs and is not counted in GDP.

Exercise 16.4

Nominal GDP is the market value of final goods and services produced in the economy. The value of final goods produced by the goods industry is \$4,000, $\$5,000 - \$1,000$ sold as intermediate inputs to the service industries. The value of final services produced by the service industries is \$9,000, $\$10,000 - \$1,000$ sold as intermediate inputs to the goods producing industries. Nominal GDP is \$13,000, i.e. $\$4,000 + \$9,000$.

The value of output is the sum of value added in the goods and service industries, namely

i. Value added in services= $\$10,000 - \$1,000$ intermediate inputs of computers, paper etc.= $\$9,000$.

ii. Value added in goods$= \$5,000 - \$1,000$ intermediate financial and other services$= \$4,000$.

iii. Value of aggregate output$= \$13,000$.

Exercise 16.5

(a) Nominal GDP by expenditures$= C + I + G + X - IM = 4,000$.

(b) Net domestic income=Employ income+Business income+invest income$= 3,850$.

(c) Nominal GDP by income=Net domestic income+capital consumption allowance+net indirect taxes$= 4,000$.

Exercise 16.6

(a) Investment expenditure is $Y - (C + G + NX) = \$2,000 - (\$1,700 + \$50 + \$40) = \$210$.

(b) If exports are \$350 and net exports are \$40 imports are \$310.

(c) Yes. Net exports would be negative if imports exceed exports.

Exercise 16.7

(a) Growth in nominal GDP from 2012 to 2013 is 10%.

(b) Real GDP in 2012 was \$721.15. Real GDP in 2013 was \$736.60. Real GDP grew by 2.14%.

(c) Per capita real GDP was \$28.8 thousand in 2012 and \$24.5 thousand in 2013.

(d) The standard of living declined because population grew faster than real GDP.

Chapter 17 Solutions

Exercise 17.1

(a) The AD and AS curves are as shown below.

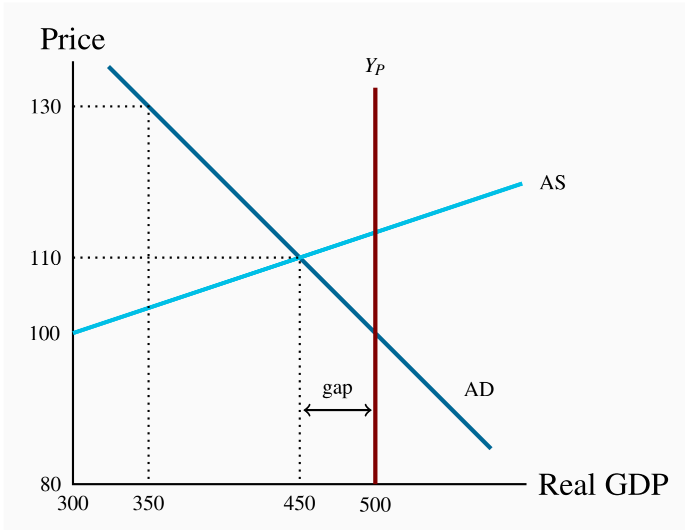
(b) The short-run equilibrium values are $P = 110$, $Y = 450$, where the AD and AS curves intersect.

Exercise 17.2

(a) The diagram shows potential output of $Y_P = 500$ added to the diagram.

(b) The diagram shows an output gap. Equilibrium $Y \neq Y_P$.

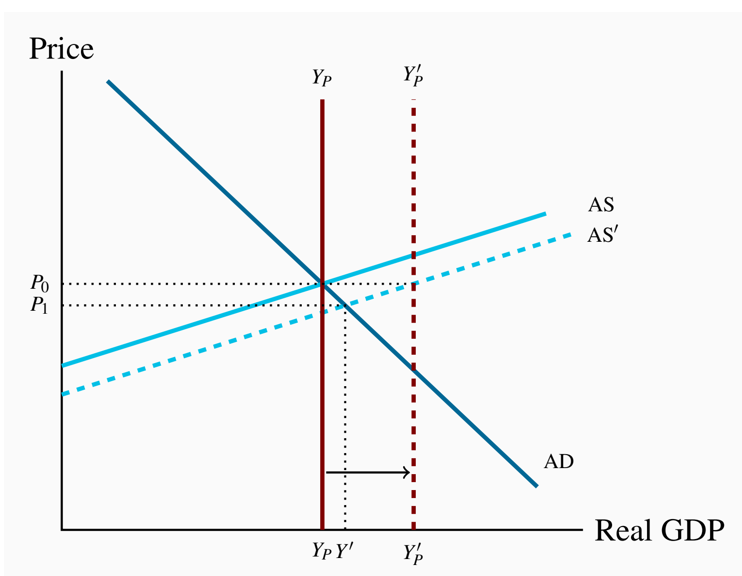
(c) The output gap is $Y - Y_P = (450 - 500) = -50$.



Exercise 17.3

Growth in labour force, or the stock of capital, or improvements in technology that increased the productivity of labour and capital would increase the economy's capacity to produce goods and services and increase potential GDP. As shown in the diagram below, the vertical line measuring potential output would shift to the right. Without an increase in AD a recessionary gap opens. The economy could produce more output at any price level without putting upward pressure on the price level beyond the initial level P_0.

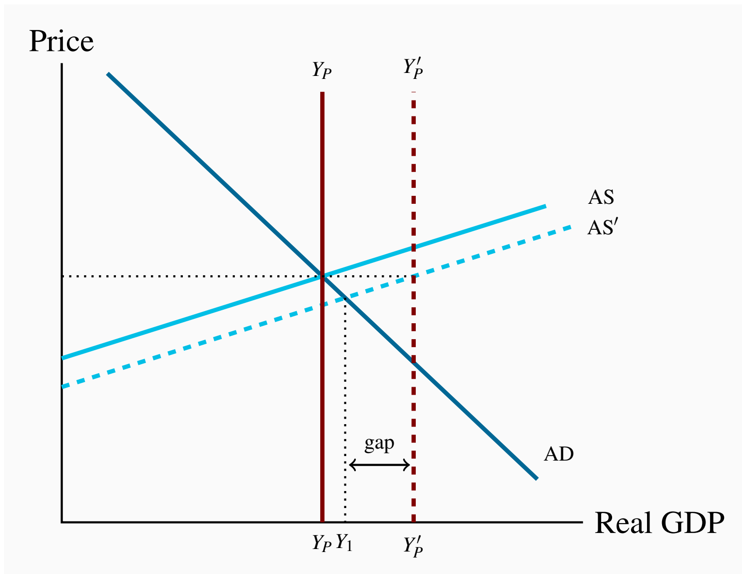
The effect of growth in labour force, or capital stock, or improved technology, is shown:



Exercise 17.4

(a) A simple approximation of the annual growth in potential output is the sum of the growth in labour force and labour productivity, namely $1.5\% + 1.0\% = 2.5\%$. However it is more accurate to recognize the compounding effect and multiply $1.015 \times 1.01 = 1.0252$, which gives an annual rate of growth of potential output as 2.52%.

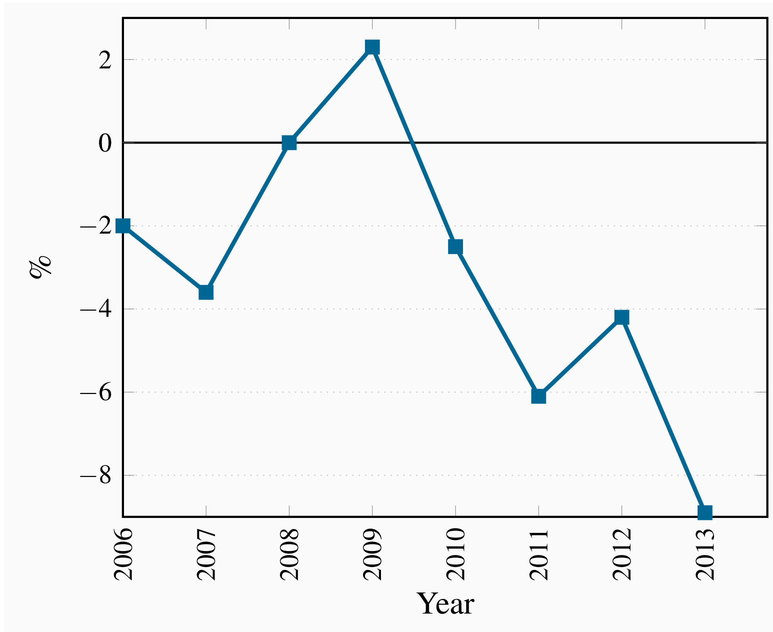
(b) The growth in potential output is illustrated by a rightward shift in the Y_P and AS curves as shown below.



Exercise 17.5

Output gaps in a growing economy are calculated as: $\frac{Y - Y_P}{Y_P} \times 100\%$. The data give the following annual gaps:

Year	2006	2007	2008	2009	2010	2011	2012	2013
Gap	−2.0	−3.6	0	+2.3	−2.5	−6.1	−4.2	−8.9

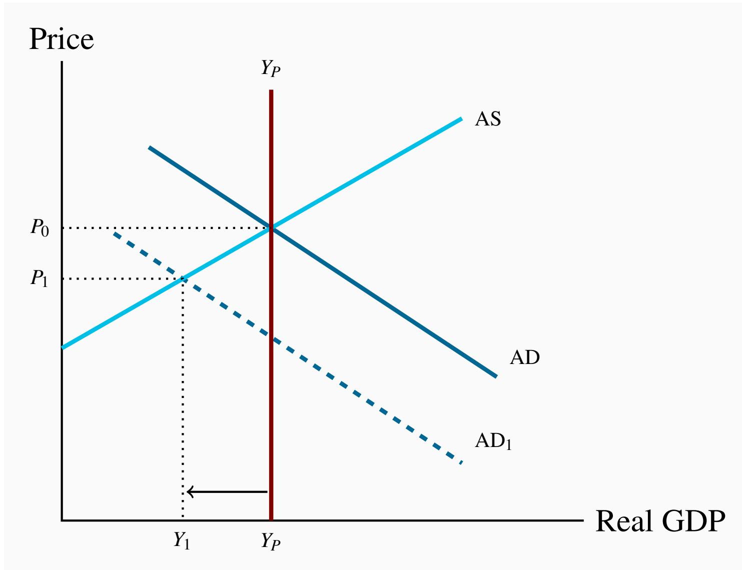


The plot of the output gaps shows the economy in recession in 2006-07, followed by recovery and boom from 2008-09, followed by recession starting in 2010, running into 2011 and then moderating in 2012 before deepening in 2013.

Exercise 17.6

(a) A slowdown in growth in Chinese imports from Westland would reduce AD in Westland from AD to AD_1 in the diagram. After this fall in demand Westland producers would reduce growth in their output and continue to reduce output reducing Y relative to Y_P as at Y_1. The price level would fall as lower of output growth reduced production costs. The economy would move toward a new equilibrium at P_1Y_1 in the diagram.

(b) Lower growth in GDP in China would lower growth in employment and raise unemployment rates in Westland if employment growth rates there were lower than labour force growth rates.



Exercise 17.7

Short-run equilibrium values for Y and P are found by solving the AD/AS model.

(a) Short-run equilibrium means AD=AS. For the aggregate demand and supply functions given:

$$\begin{aligned} Y &= 2250 - 10 \times (125 + 0.1Y) \\ 2Y &= 2250 - 1250 \\ Y &= 500 \end{aligned}$$

Then substituting $Y = 500$ into the AS function gives:

$$\begin{aligned} P &= 125 + 0.1 \times 500 \\ P &= 175 \end{aligned}$$

Short-run equilibrium values are: $P = 175$, $Y = 500$.

(b) The economy is in short-run equilibrium at potential output. The diagram is shown below.

(c) Higher oil and commodity prices increase production costs and raise the AS curve, changing the equation from:

$$\begin{aligned} \text{AS:} \quad & P = 125 + 0.1Y \ \text{ to} \\ \text{AS}': \quad & P' = 130 + 0.1Y \end{aligned}$$

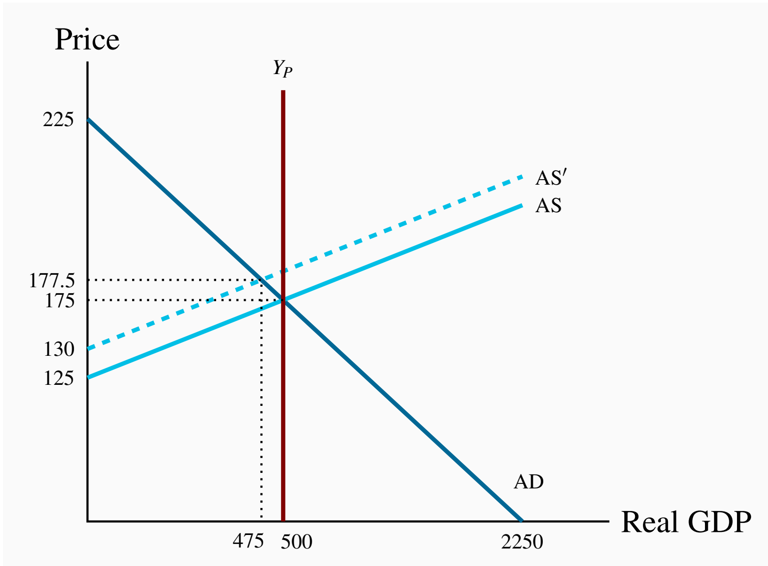
The new equilibrium Y by substituting AS$'$ in AD by:

$$\begin{aligned} Y &= 2250 - 10(130 + 0.1Y) \\ Y &= 2250 - 1300 - Y \\ 2Y &= 950 \\ Y &= 475 \end{aligned}$$

The new equilibrium price level, from AS′:

$$P' = 130 + (0.1 \times 475)$$
$$P' = 177.5$$

(d) The new AS′ curve is shown as a dotted line in the diagram. There is a recessionary gap of 25.



Chapter 18 Solutions

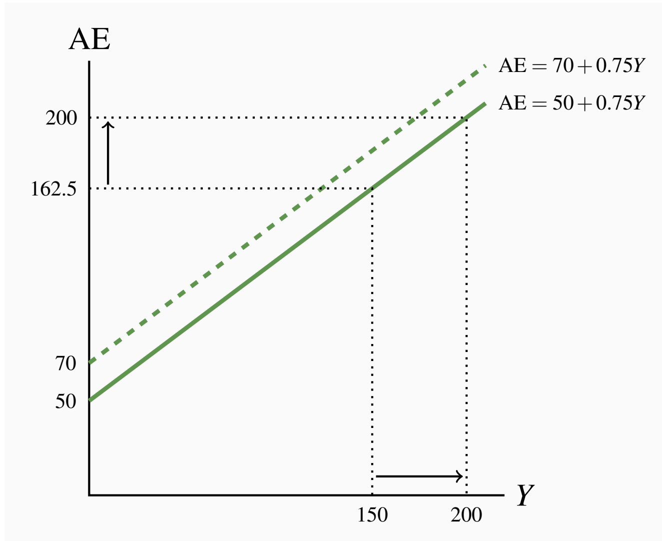
Exercise 18.1

(a) See the diagram below.

(b) If Y were to increase from 150 to 200 AE would increase by $(0.75 \times 50) = 37.5$ to 200 as shown in the diagram.

(c) Autonomous expenditure is constant at 50 in both (a) and (b) but induced expenditure increase from 112.5 when $Y = 150$ to 150 when $Y = 200$.

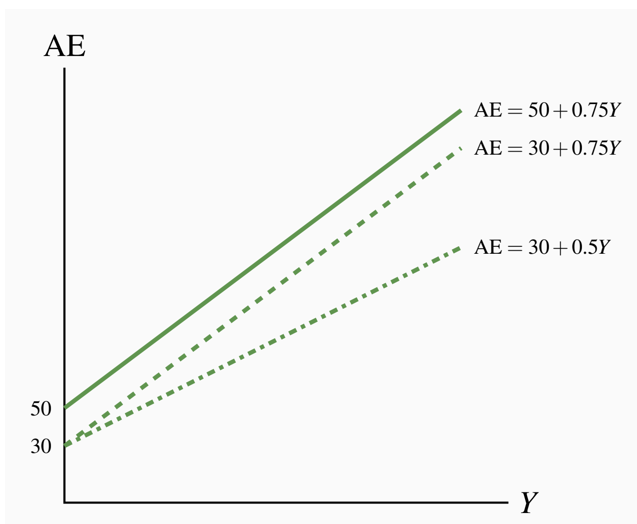
(d) If autonomous expenditure increased by 20 the AE line in the diagram would shift up as shown, with AE higher by 20 at every Y. The new AE equation would be $\text{AE} = 70 + 0.75Y$.



Exercise 18.2

(a) Initial AE: $\text{AE} = 50 + 0.75Y$. After drop in A: $\text{AE}' = 30 + 0.75Y$.

(b) After reduction in induced expenditure $\text{AE}'' = 30 + 0.5Y$ with lower slope $\Delta\text{AE}/\Delta Y = 0.5$.



Exercise 18.3

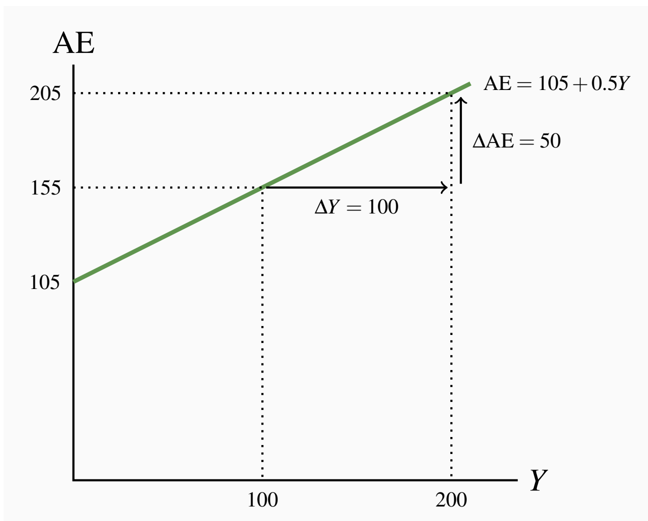
Given $\text{AE} = 105 + 0.5Y$, the table is:

Y	A	$0.5Y$	$\text{AE} = 105 + 0.5Y$
0	105	0	105
50	105	25	130
100	105	50	155
200	105	100	205

(a) In the diagram the intercept on the vertical axis is 105 and the slope of AE is 0.5.

(b) The slope of the AE function is 0.50. It measures the change in AE caused by a change in national income.

(c) The equation for AE is $\text{AE} = 105 + 0.5Y$.



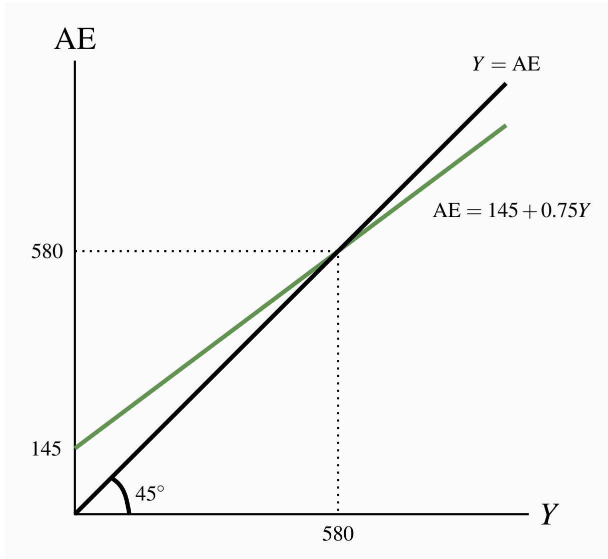
Exercise 18.4

(a) The aggregate expenditure function is $\text{AE} = C + I + X - IM$. See the diagram below.

(b) The 45° line is shown in the diagram.

(c) Algebraically, in equilibrium:

$$\begin{aligned} Y &= \text{AE} \\ Y &= 145 + 0.75Y \\ Y - 0.75Y &= 145 \\ 0.25Y &= 145 \\ Y &= 580 \end{aligned}$$



Exercise 18.5

(a) If output is OG, planned expenditure would be OB. Planned expenditure would be greater than current output.

(b) At output OG the *unplanned decrease* in inventories is AB.

(c) Business firms will respond by increasing output to meet the strong demand for output.

(d) The equilibrium level of output and expenditure is OH=OD.

(e) If output where at OJ, there would be an *unplanned increase* in inventories in the amount EF.

Exercise 18.6

(a) The initial equilibrium would be at Y =OG.

(b) If there were a decrease in induced expenditure, the slope of the AE function would decrease by a corresponding amount and the new AE function would be AJ.

(c) The new equilibrium, based on the AE function AJ, would be Y =OF.

(d) If induced expenditure increased the new AE function would be AL and the new equilibrium would be Y =OH.

Exercise 18.7

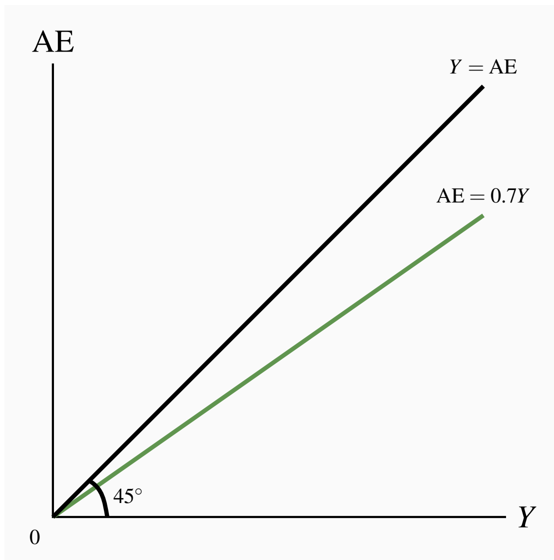
(a) When autonomous expenditure $A = 0$, induced expenditure $(c - m) = 0.7$ the AE function is:

$$\text{AE} = (c - m)Y$$
$$\text{AE} = 0.7Y$$

(b) A diagram to illustrate this AE is shown below.

(c) The equilibrium level of real GDP is $Y = 0$.

(d) The equilibrium $Y = 0$ because expenditure is not sufficient, does not provide sufficient revenues, to cover the costs of production at any positive level of real GDP.



Exercise 18.8

(a) The slope of the AE function is 0.6. This means the multiplier, the size of the change in equilibrium income as a result of a change in autonomous expenditure is $[1/(1 - \text{slope of AE})] = [1/(1 - 0.6)] = 2.5$. Then an increase in planned investment by 10 would increase equilibrium Y by $10 \times 2.5 = 25$.

(b) An increase in real GDP by 25 as a result of an increase in autonomous expenditure by 10 means induced expenditure has increased by 15.

(c) If the slope of AE were 0.8 the multiplier would be $[1/(1 - 0.8)] = 5$ and an increase in investment by 10 would increase equilibrium real GDP by $10 \times 5 = 50$. Induced expenditure would increase by 40. The larger is induced expenditure the steeper is the slope to AE and the larger the multiplier.

Exercise 18.9

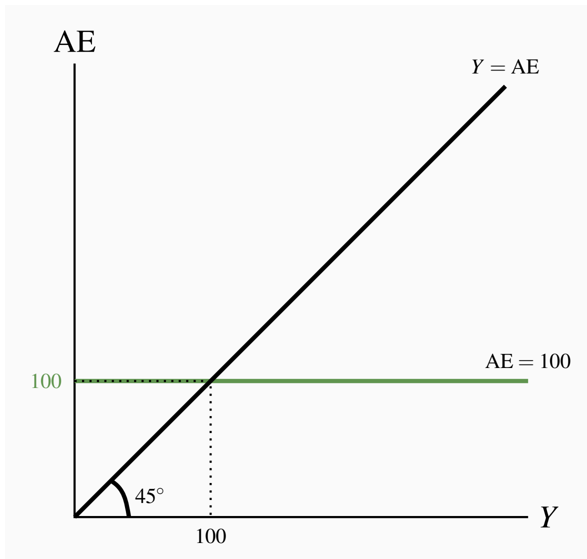
(a) The AE function would be $\text{AE} = 100$.

(b) In the diagram AE would be a horizontal line at 100.

(c) The equilibrium level of $Y = \text{AE} = 100$.

(d) If autonomous expenditure increased by 25 to 125 equilibrium in come would increase by 25.

(e) The multiplier $[1/(1 - \text{slope of AE})] = 1/1 = 1$. The slope of $\text{AE} = 0$. There is no induced expenditure to change equilibrium income by more than a change in autonomous expenditure.



Chapter 19 Solutions

Exercise 19.1

Before government is established autonomous expenditure is 200 and induced expenditure is $0.6Y$ and the aggregate expenditure function is:

$$\text{AE} = 200 + 0.6Y$$

Then for equilibrium:

$$Y = \text{AE}$$
$$Y = 200 + 0.6Y$$
$$Y_0 = \frac{200}{1 - 0.6} = 500$$

After government is established and $G = 100$: $\text{AE} = 300 + 0.6Y$. Equilibrium $Y = Y_1 = 750$.

Exercise 19.2

(a) The slope of AE in Exercise 19.1 is 0.6. The slope of AE in Exercise 19.2 is:

$$c(1-t)-m=0.75(1-0.1)-0.15=0.675-0.15=0.525$$

(b) The multiplier in Exercise 19.1 is 2.5. The multiplier in Exercise 19.2 is 2.11.

(c) With a tax rate $t=0.1$ disposable income would be $Y_d=0.9Y$ giving $\text{AE}=300+0.525Y$ and equilibrium $Y=632$, compared to the equilibrium of 750 in Exercise 19.1.

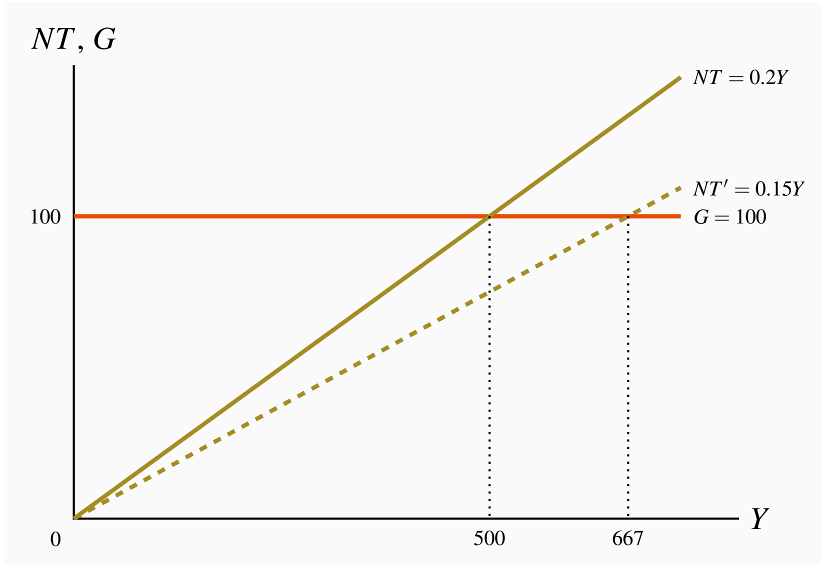
Exercise 19.3

(a) The table values are as follows:

Y	$NT=tY$	G	$BB=NT-G$
100	20	100	-80
200	40	100	-60
300	60	100	-40
400	80	100	-20
500	100	100	0
600	120	100	$+20$
700	140	100	$+40$

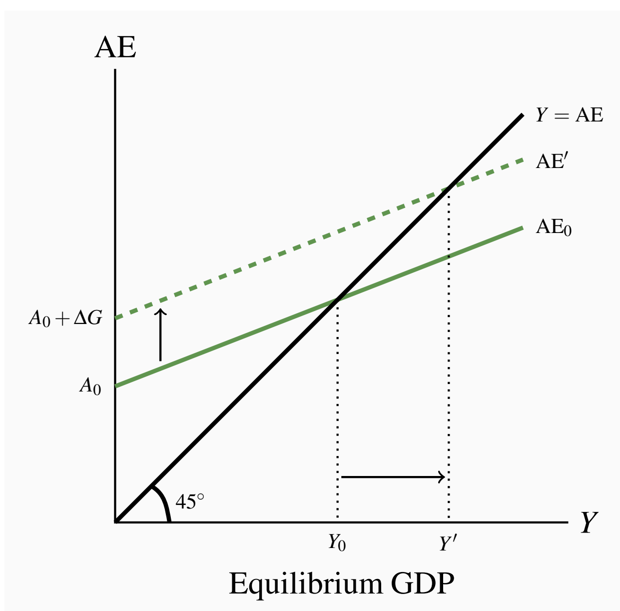
(b) The NT function has an intercept of 0 because there is no autonomous tax revenue, and a slope of 0.2. The G function has a vertical intercept at 100 and zero slope because G is autonomous. See the diagram below.

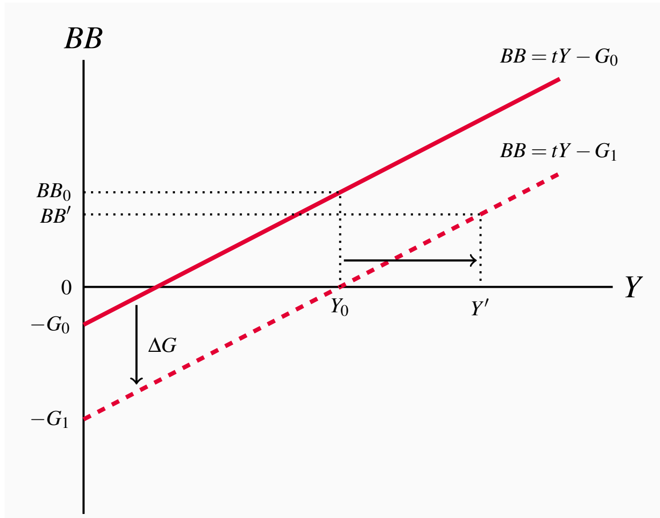
(c) A cut in the net tax rate from $t=0.20$ to $t=0.15$ reduces tax revenue proportionately at every level of Y as shown by the lower slope on the tax function NT' in the diagram. The budget balance $NT-G$ is correspondingly lower at every Y.



Exercise 19.4

Before the increase in G, equilibrium is $Y = Y_0$ and the government's budget balance is BB_0. The increase in G shifts AE up and the BB function down. The increase in Y combined with the new budget function $BB = tY - G_1$ give a new budget balance BB', smaller than the initial balance, but not reduced by the full amount of the increase in G because the expansionary effect of increased G raised Y and tax revenue.





Government budget function

Exercise 19.5

(a) The slope of AE is the change in C minus the change in IM caused by a change in real income Y.

$$\text{AE}/\Delta Y = MPC \times (1-t) - m = (0.8 \times 0.75) - 0.15 = 0.45$$

and the multiplier $\Delta Y/\Delta A = [1/(1-0.45)] = 1.82$.

(b) Autonomous expenditure is $300+400 = 700$. Equilibrium real GDP is autonomous expenditure times the multiplier:

$$Y = 700 \times 1.82 = 1{,}274$$

With equilibrium income 1,274 the government budget balance, BB, is:

$$\begin{aligned} BB &= tY - G \\ &= 0.25 \times 1{,}274 - 400 \\ &= 318.5 - 400 \\ &= -81.5 \end{aligned}$$

(c) An increase in G by 100 will increase equilibrium Y by 182 and raise tax revenue by 45.5 increasing the government budget deficit to $BB = -136$.

Exercise 19.6

(a) With $Y = 750$ and $Y_P = 850$ there is a recessionary gap$= Y - Y_P = -100$.

(b) If the $MPC = 0.75$, $MPM = 0.10$ and $t = 0.20$ the slope of $\text{AE} = 0.75(1-0.20) - 0.1 = 0.5$, and the multiplier is:

$$\frac{1}{1-0.5} = 2.0$$

To increase equilibrium income by 100 requires an increase in G by $100/2 = 50$.

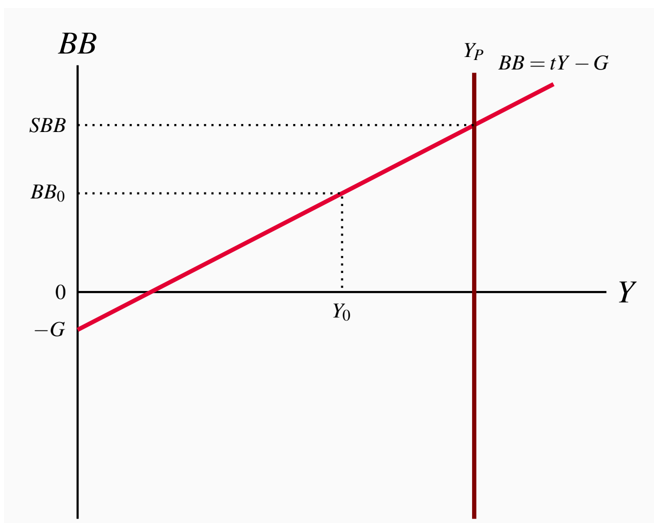
(c) With the initial tax rate $t = 0.20$, the $MPC = 0.75$ and $MPM = 0.10$ and $Y = 750$, total autonomous was $750/2 = 375$. The new tax rate needed for equilibrium $Y = 850$ and $A_0 = 375$ must give a multiplier of $850/375 = 2.267$.
Then $\frac{1}{1-0.75(1-t_1)+0.10} = 2.267$, and solving for t_1 gives $t_1 = 0.121$. As an alternative to increasing G by 50 to eliminate the GDP gap, the government could cut the net tax rate by $0.20 - 0.121 = 0.079$ or close to 40%.

Exercise 19.7

(a) The vertical intercept of BB is $-G$ and the slope of BB is the net tax rate t. See the diagram below.

(b) The structural budget balance is shown at SBB, the budget balance at Y_P. If the economy has a recessionary gap as at Y_0 the actual budget balance is BB_0.

(c) Automatic stabilization comes from the slope of the budget function. Changes in Y move the economy along the budget function, causing pro-cyclical changes in the actual budget balance but do not change the structural balance. Discretionary fiscal policy changes t, or G or both t and G. The result is a new budget function and a new structural budget balance.



Exercise 19.8

(a) With initial equilibrium $Y = 1,274 = (700 \times 1.82)$ and the government's budget balance $(NT - G) = (318.5 - 400) = -81.5$.

(b) If public debt is 500 the debt/GDP ratio is $500/1274 = 0.392$ or 39.2%.

(c) After an increase in government expenditure $\Delta G = 100$, equilibrium $Y = 1,456$ i.e. (800×1.82) and the new budget balance is $(0.25 \times 1,456) - 500 = -136$, an increase in the deficit by 54.5.

(d) In the new equilibrium the government's outstanding public debt is 636 and the debt ratio is $636/1456 = 0.437$ or 43.7%.

The deficit of 81.5 prior to the increase in G would have increased the public debt to 581.5 with no change in equilibrium income and the debt ratio to $581.5/1,274 = 0.456$ or 45.6%. The increase in G by 100 increased equilibrium income by 182, tax revenue by $182 \times 0.25 = 45.5$, the deficit by $(100 - 45.5) = 55.5$, raising the public debt from 581.5 to 637. As a result, expansionary effect of the increased government expenditure resulted in a debt ratio of $637/1456 = 43.7\%$. This is lower than the 45.6% debt ratio that would have occurred with no increase in fiscal expansion.

Chapter 20 Solutions

Exercise 20.1

Money is anything generally accepted as a means of payment, a store of wealth and a unit of account.

In Canada today, Bank of Canada notes, coins and bank deposits are money.

The money supply is the sum of notes and coin in circulation outside the banking system, and bank deposits.

A debit card is money because, like a cheque, it transfers bank deposits from the bank account of the payer to the bank account of the payee. There is no payment further financial obligation. Using a credit card to make a payment creates a credit card debt for the payer that must be settled by a money payment – notes or bank deposits.

Exercise 20.2

A central bank, like the Bank of Canada works to control money and financial conditions in the economy using its position as monopoly supplier of monetary base – bank reserve assets. It is not profit oriented. It does not attempt to make a profit.

A commercial bank works to earn a profit for its owners (shareholders) by providing banking services to the non-bank public on terms that generate net interest income. It is profit oriented.

Exercise 20.3

Banks create money by issuing their own deposit liabilities (IOUs) in payment for the assets they buy, such as financial securities and customer loan contracts.

Suppose banks operate to a 5% reserve ratio. $rr = 0.05$.

The following balance sheets show the initial new deposit and the deposit creation that follows.

(a) The new deposit provides the banks with a 100 increase in cash (reserve asset) in exchange for 100 in new deposit liabilities.

(b) In part (a) banks hold \$95 excess reserves based on the reserve ratio of 5%. In part (b) they make loans equal to their excess reserves \$95, and pay for those loans by creating new deposit liabilities, \$95.

(c) Assuming the public uses bank deposits as money and does not withdraw cash from the banking system, the banks expand their lending and create new deposits to a total of $+1,900$, based on the initial increase in cash reserves and the reserve ratio of 5%. The deposit multiplier is: $\Delta D = \Delta R/rr = 100/0.05 = 2,000$.

Table for part (a):

All banks			
Assets		Liabilities	
Cash	$+100$	Deposits	$+100$

(excess reserves $+95$)

Table for part (b):

Assets		Liabilities	
Loans	$+95$	Deposits	$+95$

Table for part (c):

Assets		Liabilities	
Cash	$+100$	Deposits	$+2,000$
Loans	$\underline{+1,900}$		
	$+2,000$		

Exercise 20.4

If banks have a reserve ratio of $rr = 10\%$ a new cash deposit of \$1,000 to the banking system would allow an expansion of bank deposits by:

$$\begin{aligned}\Delta D &= \Delta MB \times [1/(rr)]\\ \Delta D &= 1{,}000 \times [1/(0.10)]\\ \Delta D &= 10{,}000\end{aligned}$$

Deposit expansion beyond the initial \$1,000 would be the result of a \$9,000 increase in bank lending.

Yes. If the banks could encourage the public to hold less cash the banks have a larger share of the monetary base. These larger reserves would support more bank lending and deposit creation.

Exercise 20.5

Confidence in the banking system is based partly on the established reputations of banks in converting deposits into cash and the general acceptability of bank deposits as means of payment. This confidence is reinforced by the insurance coverage on deposits up to \$100,000 provided by CDIC.

Exercise 20.6

The money multiplier is defined as the change in money supply that results from a change in monetary base, or $\Delta M_S/\Delta MB$. For example, a monetary base of 1,100, of which the public holds 100 in cash and a reserve ratio of $rr = 0.10$, the money supply function is:

$$M_S = \left(\frac{1}{rr} \times MB - 100\right) + 100 = \left(\frac{1}{0.10} \times 1{,}000\right) + 100 = 10{,}100$$

If the monetary base decreased by 100, money supply would decrease by:

$$\Delta M_S = \frac{1}{rr} \times \Delta MB = \frac{1}{0.10} \times (-100) = -1{,}000$$

In this example the money multiplier is: $\Delta M_S/\Delta MB = (-1{,}000)/(-100) = 10 = 1/rr$.

Exercise 20.7

The financial crisis made both banks and the non-bank public more concerned about the risks attached to making loans and to holding bank deposits. The banks responded by increasing their reserve ratios and by being more selective about the quality of loans and other assets they bought. Even if total bank assets were not reduced overall some forms of bank credit did dry up and some potential borrowers are denied credit.

Concerns about the stability of banks and other financial institutions led the non-bank public to hold more of their money in cash rather than deposits. If the share of the monetary base held by the public as cash is increased and the banking system's lending capacity is reduced.

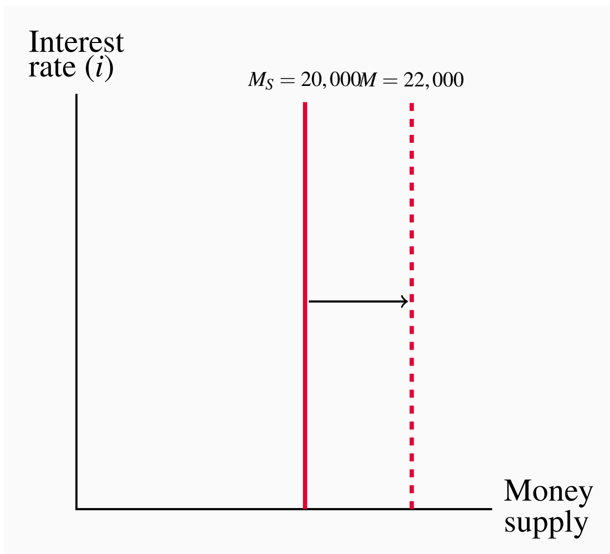
The effects of increased reserve ratios are smaller deposit multipliers. Higher public cash holdings reduce bank reserves, reduce bank lending and reduce money supply in the absence of an offsetting increase in the monetary base.

Exercise 20.8

With monetary base $MB = \$1{,}000$, $rr = 0.05$ and assuming no cash held by the public the money supply would be:

$$M_S = 1{,}000 \times \left(\frac{1}{0.05}\right) = 1{,}000 \times 20 = 20{,}000$$

This money supply is shown by the vertical line M_S in the diagram. Changes in interest rates do not affect the size of the money supply in our simple examples.



If the monetary base were to increase by 10 percent to 1,100 the money supply would be increased by $(100 \times (1/0.05) = 2{,}000$ to 22,000. The money supply function in the diagram is shifted to the right to show this effect.

Chapter 21 Solutions

Exercise 21.1

A perpetual bond with a 3% coupon would have a market price of \$100 if the current market rate were 3%. (Yield = coupon/ price. Price = coupon/yield = \$3.00/0.03 = \$100)

If current market rates fell to 2.5% you would be pleased. The price of your 3% perpetual bond would be $\$3.00/0.025 = \$120$. You would have a 20% capital gain if you sold your bond.

Exercise 21.2

(a) The market price of the bond is the present value of the future stream of payments it provides. A two-year 5% bond, when market rates are 5% has a present value of:

$$P_B = PV = \frac{\$5}{1.05} + \frac{\$105}{1.05^2} = \$4.76 + \$95.24 = \$100$$

(b) If market rates rise to 6%, then the market price of the two-year 5% bond will be:

$$P_B = PV = \frac{\$5}{1.06} + \frac{\$105}{1.06^2} = \$4.72 + \$93.45 = \$98.17$$

A rise in market interest rates reduces the prices of outstanding bonds.

(c) The market risk in holding bonds is that bond prices and interest rates vary inversely. A rise in market interest rates means a fall in bond prices and capital losses for bond holders, at least on paper. A fall in market rate means a rise in bond prices.

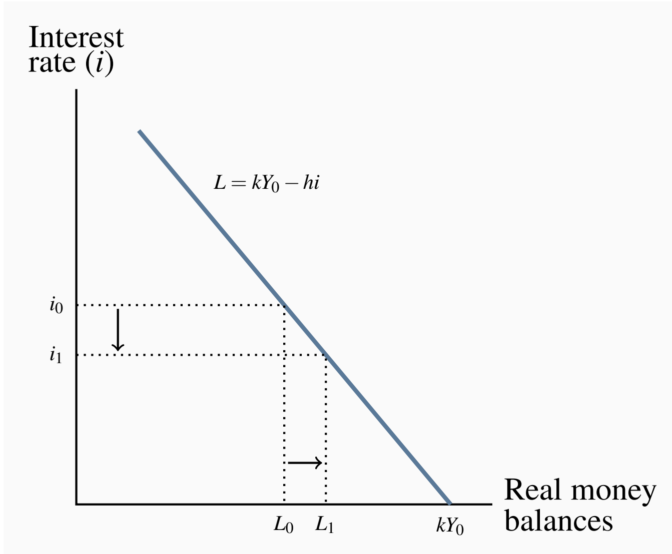
Exercise 21.3

(a) The horizontal intercept in the diagram (below) shows the money balances people would want to hold, based on their income (Y_0) if the interest rate, and thus the opportunity cost of holding money were zero. The slope of the line shows how portfolio managers would adjust their holdings of money (vs bonds) if interest rates were to change.

(b) At the interest rate i_0 the demand for money balances is L_0. Some money balances are held to make regular payments, some to provide for uncertainty in the timing of receipts and payments, and some to lower the risks in portfolios of bonds and money.

(c) If interest rates dropped from i_0 to i_1 in the diagram, the people's demand for money balances would increase from L_0 to L_1. Lower interest rates mean lower opportunity costs to holding money balances.
A fall in interest rates increases the demand for money balances and reduces the demand for bonds. Portfolios shift from bonds to money.

(d) An increase in real GDP ($\Delta Y > 0$) would shift the demand for money function to the right by the amount $k\Delta Y$. Higher Y means higher income and expenditure levels and a need for larger transaction balances to make payments.



Exercise 21.4

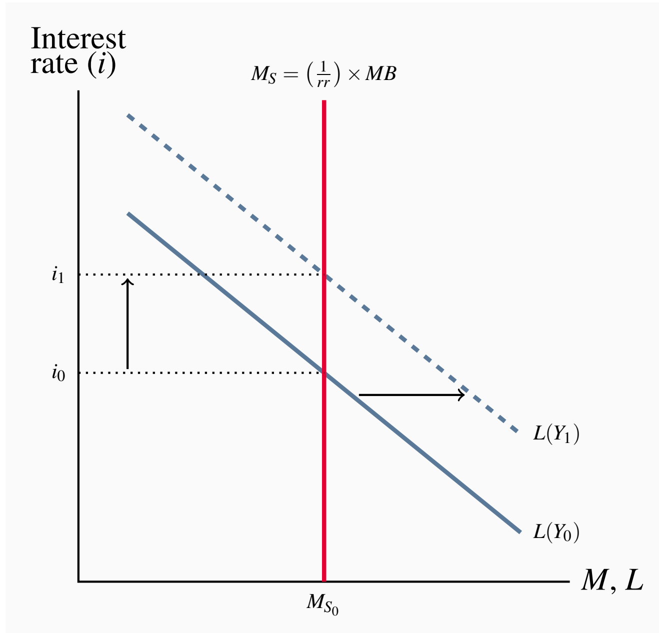
If US interest rates rose the Canadian dollar price of a $US would rise. A rise in US interest rates relative to Canadian rates increases the returns to holding US$ bonds relative to Cdn$ bonds. Portfolio managers would demand more US$ in order to increase holdings to US$ bonds relative to Canadian bonds. Increased demand for US$ raises the Cdn$ price of the US$ to more than $1.25Cdn. The Canadian dollar depreciates.

Exercise 21.5

(a) In the diagram the M_S line is vertical. M_S is not affected by interest rates. The monetary base and the money multiplier determine the position of the M_S line. The position of the demand for money function L is determined by the level of real income Y_0, and the transactions demand kY_0. The slope of the L function illustrates the reaction of portfolio managers, changing the mix of money and bonds in portfolios in response to changes in interest rates. Higher interest rates reduce the demand for money balances and increase the demand for bonds.

(b) An increase in income to Y_1 shifts the demand for money function L to the right to $L(Y_1)$. The excess demand for money balances at the initial interest rate i_0 results in the sale of bonds, bond prices fall and yields increase until interest rates rise to i_1.

(c) With higher Y and higher demand for money the new equilibrium interest rate is higher but money holdings are unchanged because the money supply is fixed.

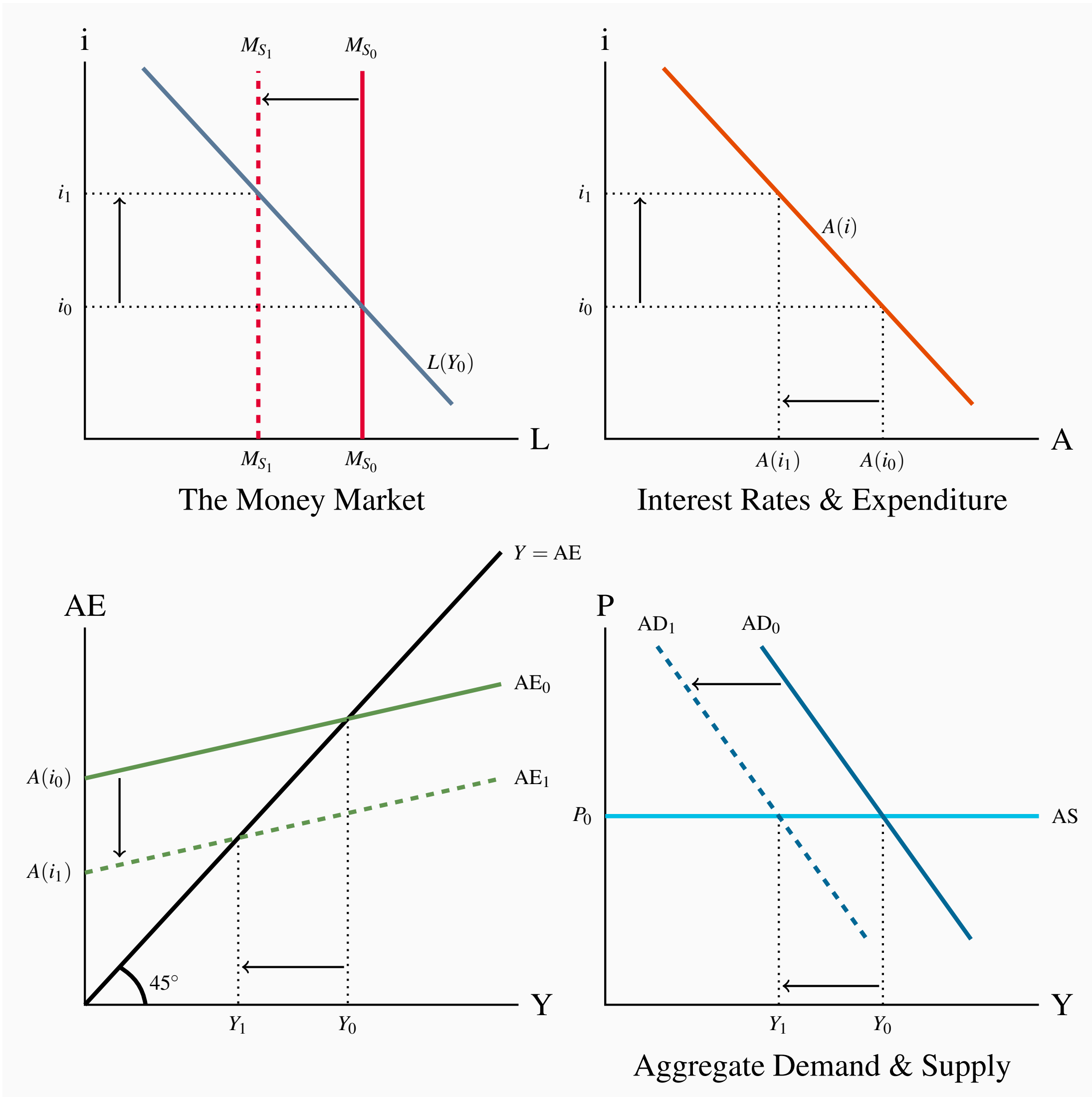


Exercise 21.6

(a) A reduction in M_S shifts the M_S line to the left, raising interest rates and lowering expenditure to $A(i_1)$. AD shifts to the left by the change in A times the multiplier to AD_1 and Y is reduced to Y_1.

(b) Alternatively, an increase in precautionary demand for money would shift the demand for money (L) to the right, raising interest rates and lowering expenditure and aggregate demand and equilibrium Y.

(c) Alternatively, any increase in autonomous expenditure would shift the aggregate expenditure function up, the AD function to the right and higher Y would increase the demand for money balances L. As a result interest rates would rise offsetting some of the increase in investment expenditure (moving along the $A(i)$ function) and shifting AD back to the left. The net result would be a increase in investment, interest rates and equilibrium Y.



Chapter 22 Solutions

Exercise 22.1

A central bank that operated to make maximum profits would cause financial market instability by expanding the monetary base through its purchase of interest bearing government bonds until the yields on those bonds were driven to (approximately) zero.

A commercial bank pursues profits as long as the costs of raising funds through deposit expansion are less than the interest revenue earned by expanding its lending. The bank's shareholders expect a positive return on their equity in the business. Competition among banks and the public's concern about the solvency of a bank means the costs of funds rises as the bank expands, thereby eliminating the profitability of further expansion.

Currently, the central bank's operating objective is to control the rate of inflation based on an agreed target inflation rate of 2%.

The central bank's unique position as monopoly supplier of the monetary base, cash and central bank deposits, gives it the power to pursue its monetary policy objectives.

Exercise 22.2

Monetary base, central bank notes (cash) and deposits, are the ultimate means of payment in the economy. Commercial banks issue deposits that are convertible into cash on demand. This convertibility together with the public's demand for cash balances creates a demand for monetary base that limits the size of commercial bank deposit liabilities. The central bank's control of the monetary base gives it control of the money supply and interest rates.

Exercise 22.3

A change in monetary base is by itself a direct change in the money supply. But the profit seeking behaviour of the commercial banks causes a larger change in the money supply as a result of ΔMB. If $\Delta MB > 0$, for example, and the public deposits these new funds in the banks, the banks find they are holding excess reserves. These reserves support an increase in bank lending and the deposit creation that goes with it. The money supply increases based on the money supply multiplier provided the currency and reserve ratios are constant.

(a) The purchase of \$10 million in the open market by the central bank creates \$10 million in monetary base, which increase the reserves of the commercial banks, provided it is not held as cash by the non-bank public.

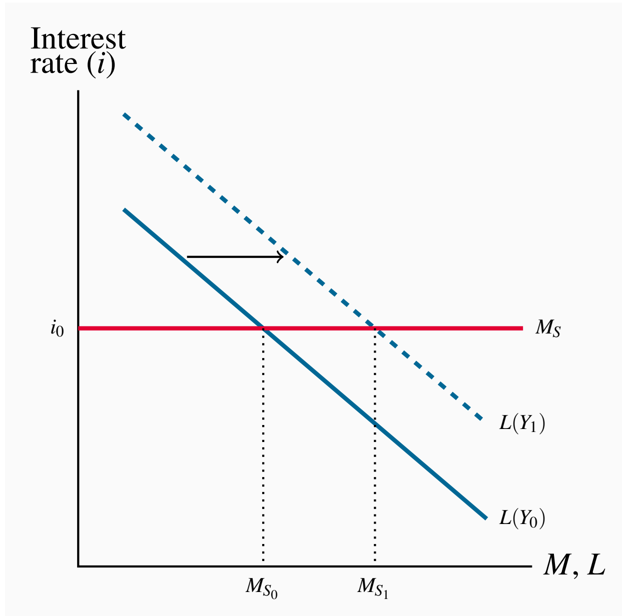
(b) With a reserve ratio $rr = 0.025$ a \$10 million increase in monetary base results in:
   i. An increase in money supply of \$400 million. The money supply multiplier is $[(1/0.025) = 40]$
   ii. An increase in bank reserve balances of \$10 million, the full amount of the increase in the monetary base because the public does not increase cash holdings.

Exercise 22.4

To set and maintain interest rate at i_0 in the diagram the central bank provides whatever money supply is demanded at that interest rate. This is shown in the diagram by the horizontal M_S line at i_0.

An increase in real output from Y_0 to Y_1 would increase the demand for money, shifting the $L(Y)$

line in the diagram to the right to $L(Y_1)$. The money supply would increase to M_{S1} to meet the increased demand for money at the interest rate i_0.



Exercise 22.5

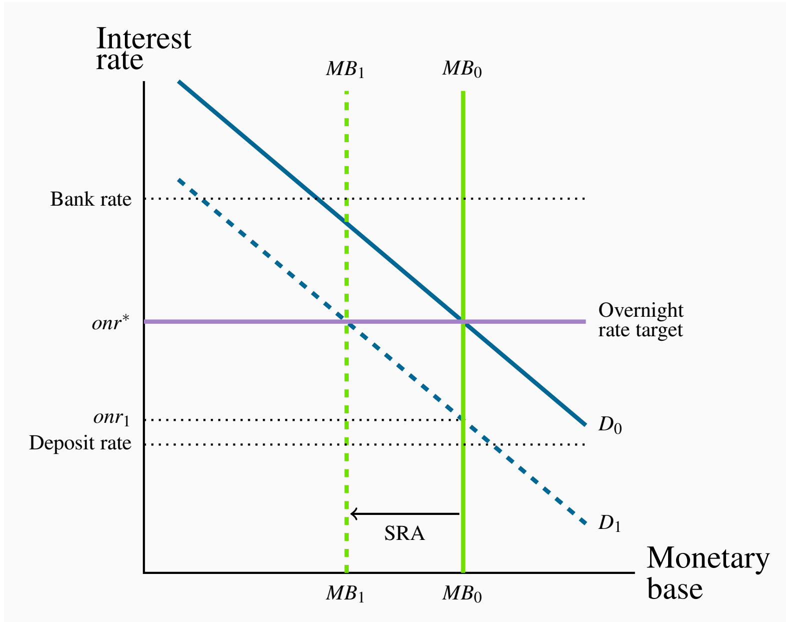
(a) The Bank of Canada's monetary policy target is an inflation control target of 2% in the CPI within a range of 1%-3%. The Bank aims at a 2% inflation rate over six to eight quarters.

(b) The Bank uses the overnight interest rate as its monetary policy instrument to influence short-term interest rates by raising or lowering the target and operating band it sets for the overnight rate.

(c) Implementing monetary policy by setting the interest rate means the Bank gives up its short-term control of the money supply.

Exercise 22.6

The market for overnight funds:



The Bank has set its target for the overnight rate at onr* consistent with a demand for monetary base D_0. If it happens that the demand for monetary base is less than D_0, for example D_1, the ONR will fall below the Bank's target. To prevent this the Bank can remove some monetary base from the overnight market by selling short term securities to the banks on the agreement that it will buy those securities back the next day. This is a 'sale and repurchase agreement', and SRA. Its effect is to reduce the monetary base, as illustrated by the shift to MB_1, and maintain onr* as shown by the intersection of D_1 and MB_1.

The Bank uses an SRA rather than an open market operation because an SRA makes an immediate change in the clearing balance position of the banks that last for one day. It is a very short-term adjustment. An open market operation takes time to affect the clearing balance position of the banks as it works its way through bond markets and bank customer accounts to clearing balances. It is better suited to longer term management of the monetary base rather than very short term and likely temporary adjustments to bank clearing balances.

Exercise 22.7

(a) The central bank chooses and sets the interest rate it thinks will be consistent with equilibrium at potential output at its target rate of inflation.

(b) A rise in the unemployment rate would indicate a fall in aggregate demand and call for a decrease in the Bank's interest rate from its basic setting.

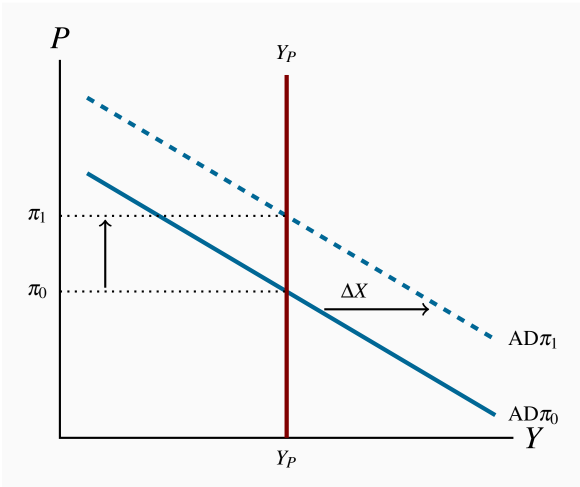
(c) An inflation rate above the Bank's target would indicate stronger than expected aggregate demand and call for a rise in the Bank's interest rate from its basic setting.

(d) A persistent change in either unemployment or inflation would lead the Bank to change its basic interest rate setting.

Chapter 23 Solutions

Exercise 23.1

The equilibrium inflation rate is determined by AD and Y_P at the expected inflation π^e. An increase in export demand would shift AD to the right and raise the equilibrium inflation rate and real Y.



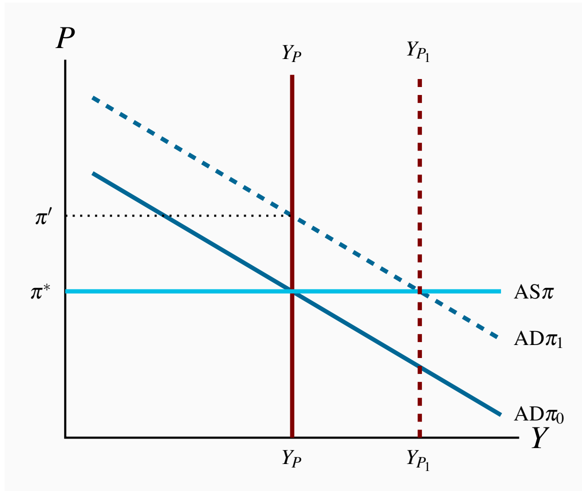
If the central bank reacted to defend its inflation target as π_0, it would raise its interest rate and money supply growth would decrease to shift AD back from AD_1 to AD_0.

Exercise 23.2

If the central bank reacted to defend its inflation target as π_0, it would raise its policy interest rate and money supply growth would decrease to shift AD back to an intersection with AD at $\pi_0 Y_P$.

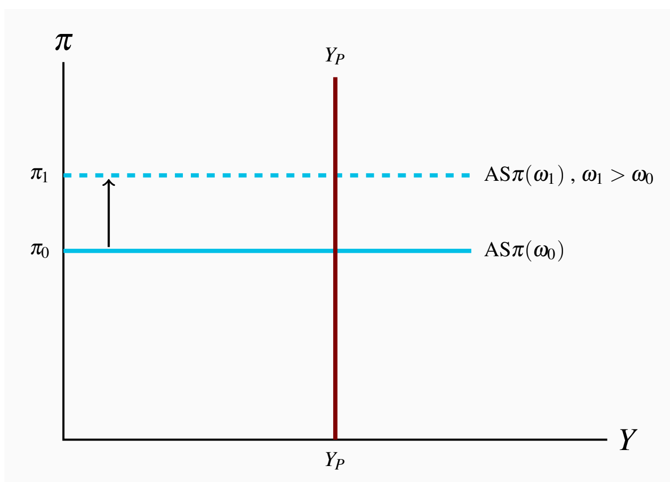
Exercise 23.3

Increased investment shifts AD to the right in the AD/AS diagram to AD_1, increasing real GDP and putting upward pressure on the inflation rate. However, as the new capital stock and technology comes on stream Y_P grows, shifting the long run AS right to Y_{P1} and offsetting the initial inflationary gap. The final result is a higher level of real GDP at Y_{P1} with inflation at the initial level π_0, the central bank's target.



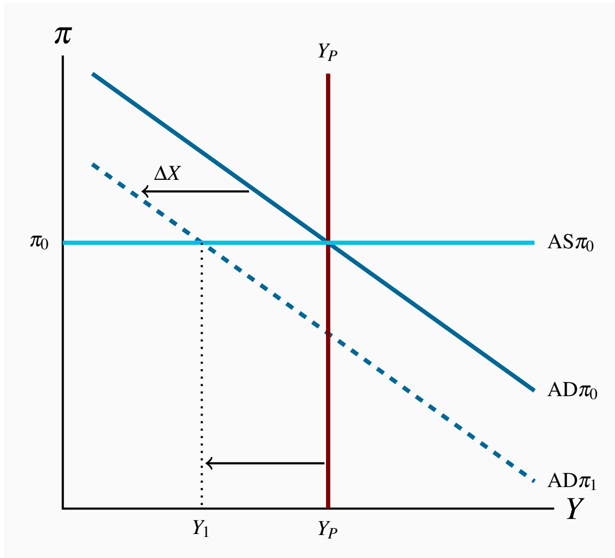
Exercise 23.4

A higher rate of increase in wages shifts AS up as the rate of increase in real costs of production rises at each level of output.

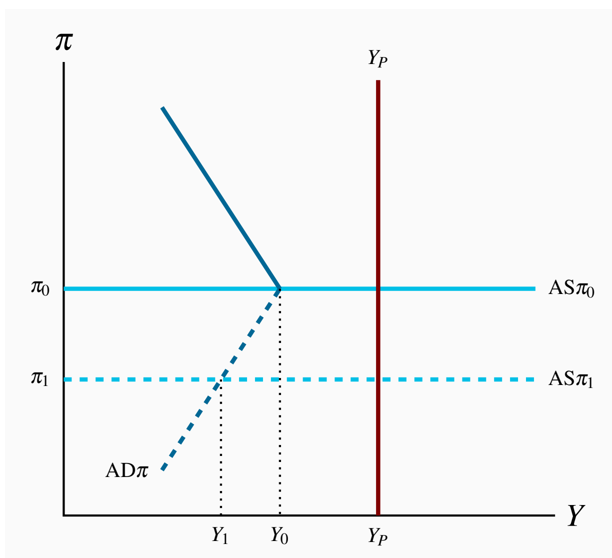


Exercise 23.5

The economy is initially in equilibrium at Y_P and π^*. A fall in exports shift ADπ left to ADπ_1, opening a recessionary gap $Y_1 - Y_P$. The inflation rate is little changed in the short run, moving left along the AS curve but will decline over time if the recessionary gap persists, reducing rates of increase in money wage rates and shifting AS down.



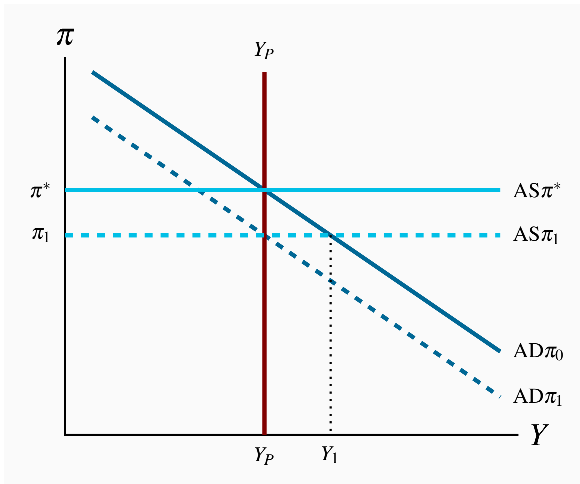
Exercise 23.6



With the economy at Y_0 and the central bank having reduced its policy rate to the lower bound to fight the persistent recessionary gap, the ADπ curve has a positive slope at any inflation rate less than $\pi = 0$. Any fall in the inflation rate from π_0 raises the real interest rate $(i - \pi)$ because the bank cannot counter with a lower nominal interest rate. Higher real rates reduce expenditures and output.

Furthermore, cuts in the rate of increase in money wage rates or any other cost reductions that lower the inflation rate that shift ASπ down, will increase rather than reduce the recessionary gap: e.g. $Y_1 - Y_P$ in the diagram.

Exercise 23.7



In an $AD_\pi/AS_\pi/Y_P$ diagram the economy is at π^*, Y_p before the tax cuts. Cutting the GST lowers $AS\pi$ to $AS\pi_1$ and the economy moves along $AD\pi$. The fall in the inflation rate and its effect on aggregate expenditure lead to an inflationary gap $Y_1 - Y_p$.

The Bank will not react if it sees the cut in the GST as a short term *transitory* disturbance that will affect the price level and inflation rate for about one year. It will leave its policy rate unchanged. With underlying AS conditions unchanged, $AS\pi$ will return to $AS\pi^*$, the inflationary gap will disappear and inflation rate will rise to π^*. Households will benefit from a lower rate of taxation.

Alternatively, if the Bank sees the cut in the GST as one in a series of cuts it will be concerned about a persistent shift in $AS\pi$. Then the inflationary gap is a threat to the Bank's target inflation rate π^*. The Bank reacts by raising its policy interest rate to reduce $AD\pi_0$ to $AD\pi_1$ and restore $Y = Y_p$, the inflation rate is then π_1, below its target, for as long as the price level and inflation rate effects of the GST cut last and then returns to π^* as $AS\pi$ returns to $AS\pi^*$. Households and businesses enjoy lower rates of GST on expenditures but face higher interest rate costs on debt.

Exercise 23.8

The public debt is the total dollar value of government bonds outstanding. That total is the cumulative sum of federal government budget balances in Canada since Confederation in 1867. The public debt would increase in any year the government budget was in deficit or decrease in any year in which there was a budget surplus.

$$\Delta PD = -PBB + iPD$$

Exercise 23.9

The public debt ratio is PD/Y. The interest rate on the public debt, i_{PD}, increases the public debt each year by $i_{PD} \times PD$ to $(1+i_{PD})PD_0$. If the rate of growth of nominal GDP is g_Y then Y increases each year by $(1+g_Y)Y$. Then if the government's budget balance is zero, adding no new debt, the annual change in the debt ratio is $(1+i_{PD})/(1+g_Y)$. With $i_{PD} < g_Y$ and the primary budget is balanced, the public debt ratio declines as GDP grows faster than the debt. Even with a small primary budget deficit, the debt ratio *could* decline as long as $g_Y > i_{PD}$.

Exercise 23.10

A surplus in the PBB/Y reduces ADπ and reduces rate of growth of nominal GDP, $\Delta Y/Y$. The fiscal austerity coming from the primary budget surplus may lower $\Delta Y/Y$ relative to i to an extent that the debt ratio rises despite the budget surplus. This has been the recent experience of several European countries.

Because the debt ratio is the ratio of debt to income its evolution is driven by differences between the growth in the debt (the numerator) relative to growth in nominal GDP (the denominator). Debt ratio control and/or reduction through changes in the primary budget balance also affect the rate of growth of nominal GDP.

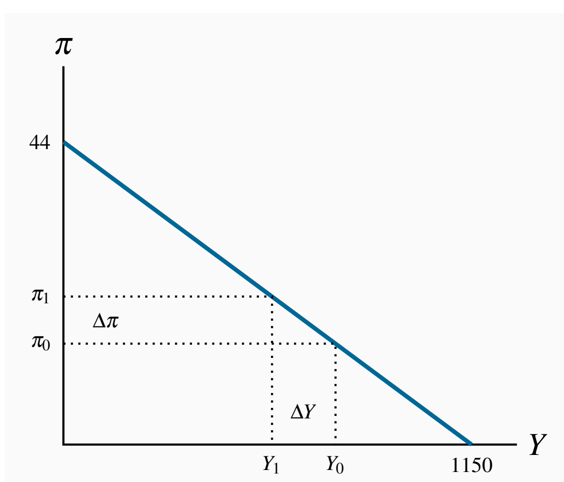
Exercise 23.11

(a) The AE function in this case is:

$$\text{AE} = C + I + G + NX$$
$$\text{AE} = 435 + 0.6Y - 5i$$

Monetary policy sets the interest rate according to $i = 3.0 + 2.0(\pi - \pi^*)$ and with an inflation target $\pi^* = 4.0$, $i = -5.0 + 2\pi$. The equation for the AD curve is then $Y = 1,150 - 25\pi$.

(b) The AD curve has a horizontal intercept at $1,150$ with slope $\Delta\pi/\Delta Y = -25$.



Exercise 23.12

With AD: $Y = 1,150 - 25\pi$ and $Y_P = 1,000$ the equilibrium inflation rate is:

$$\begin{aligned} 1,000 &= 1,150 - 25\pi \\ 25\pi &= 150 \\ \pi &= 6.0 \end{aligned}$$

This inflation rate is above the central bank's $\pi^* = 4.0$ target. Defending the inflation rate target of 4.0 calls for the central bank to raise its interest rate to:

$$i = 3.0 + 2.0(6.0 - 4.0) = 7.0$$

Chapter 24 Solutions

Exercise 24.1

(a) The balance of payments is always zero, neither deficit nor surplus. Any deficit or surplus on current account that is not offset by a surplus or deficit on financial account results in a change in official reserve holdings, and a corresponding entry to balance the balance of payments account. The current account balance, the financial account balance and the change in official reserves sum to zero.

(b) Official reserves would increase by $2 billion, the difference between the surplus on current account and the deficit on financial account.

(c) The central bank would buy foreign currency to add to the official reserve account.

(d) The monetary base increases. The central bank pays for the foreign currency it buys by issuing new central bank deposits, which are monetary base.

Exercise 24.2

The US experienced the higher inflation rate as prices doubled over 10 years while Canadian prices increase by 75%. A nominal exchange rate of $1.05 Cdn for $1.00US would preserve the real exchange rate. The Canadian dollar appreciated in terms of US dollars and the US dollar depreciated in terms of Canadian dollars.

Exercise 24.3

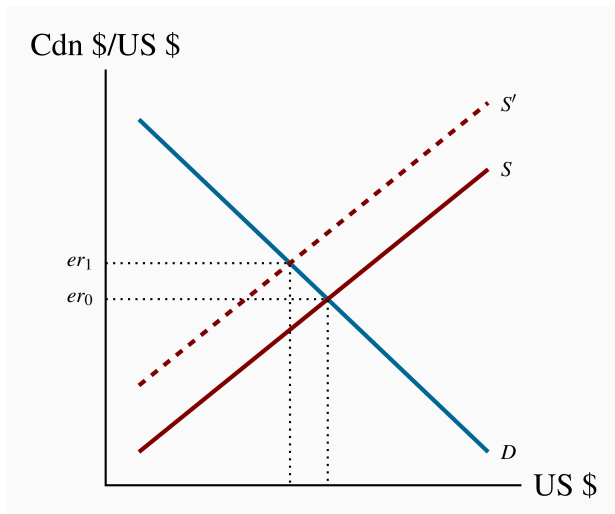
The purchase of US government securities by Canadian portfolio managers is an import of securities. Payment for these securities is made to residents of other countries and capital flows out from Canada to other countries. The financial account balance in the balance of payments is reduced.

Exercise 24.4

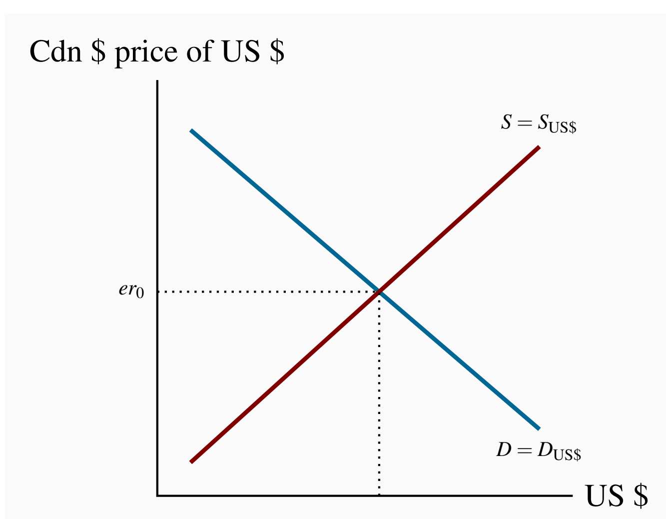
A nominal interest rate in Canada that is higher by 1 percent than the nominal interest rate in the US means the expected rate of appreciation the US dollar, relative to the Canadian dollar is 1 percent if interest rate parity prevails.

Exercise 24.5

A sharp and persistent drop in natural gas and crude oil prices would lower Canadian export receipts and the supply of US dollars on the foreign exchange market. The Canadian exchange rate would rise as the Canadian dollar depreciated.



Exercise 24.6

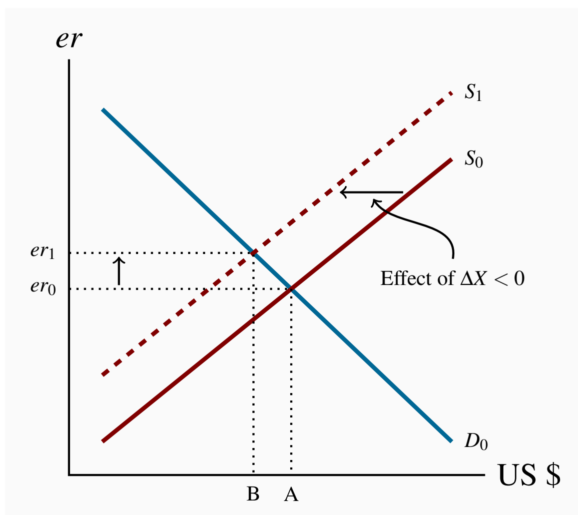


(a) The demand for foreign exchange (US dollars) in the diagram comes from Canadian demand for imports of foreign goods and services on the current account and foreign financial assets on the financial account. The demand for goods and services comes from Canadian incomes and tastes in terms of the propensity to import and the real exchange rate. The demand for foreign assets comes from Canadian portfolio decisions to hold foreign assets based on interest rate differentials and expected returns.

(b) The supply of foreign exchange (US dollars) comes from foreign demand for Canadian goods, services and assets based on foreign incomes and tastes, the real exchange rate, interest rate differentials and expected returns.

(c) The equilibrium exchange rate er_0 is the exchange rate at which the balances on current account and financial account sum to zero.

Exercise 24.7

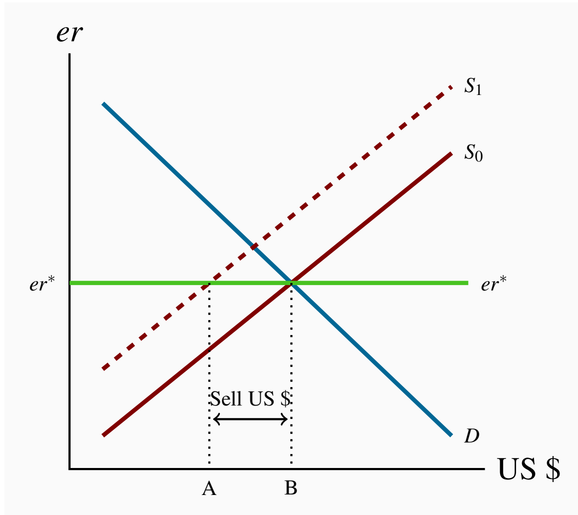


(a) A decline in exports reduces the supply of foreign exchange on the foreign exchange market, shifting the supply curve to the left from S_0 to S_1 in the diagram. The exchange rate increase from er_0 to er_1 as the domestic currency depreciates.

(b) The rise in the exchange rate raises the domestic currency price of imports while at the same time making exports more price competitive in foreign markets and more profitable for domestic producers. Expenditures on imports decline while export revenues rise which maintains balance of payments equilibrium.

(c) The adjustment to the initial fall in exports comes from the flexibility and change in the foreign exchange rate. There is no government intervention and no change in official reserve holdings.

Exercise 24.8



(a) Fixed exchange rate (er^*):

(b) The decline in exports reduces the balance on the current account in the balance of payments and reduces the supply of foreign exchange in the foreign exchange market. In the diagram the decline in the current account balance, measured in US dollars is $U_0 - U_1$, $(= er^* \times (U_0 - U_1)$ in Cdn \$). The supply curve shifts to the left by this amount.

(c) To defend the fixed exchange rate at er^* the central bank sells foreign exchange from the official reserve account equal to the AB in the diagram, the difference between market supply and demand at the fixed rate er^*.

(d) The central bank sale of US $ would reduce holdings of official reserves and the monetary base.

Exercise 24.9

Flexible exchange rates provide for two linkages in the transmission mechanism for monetary policy. Changes in interest rates in pursuit of short-term stabilization objectives produce complementary changes in the exchange rate. Higher interest rates cause an appreciation of the domestic currency and lower interest rates depreciation. Monetary policy has simultaneous effects on domestic expenditures and net exports. By contrast, with flexible exchange rates and a monetary policy that controls money supply, fiscal policy is weakened by both interest rate and exchange rate crowding out. Fiscal expansion raises income and the demand for money pushing interest rates up and lowering the exchange rate with the result that investment and net exports are reduced.

Exercise 24.10

With a fixed exchange rate policy, interest rates must be maintained at the level required by the fixed exchange rate. Expansionary fiscal policies that raise real GDP and the demand for money must be matched by an expansion in the money supply to keep interest rates from rising. As a result crowding out does not impair the power of fiscal policy as it would in a closed economy or in an open economy with flexible exchange rates.

Chapter 25 Solutions

Exercise 25.1

(a) Growth in potential GDP is growth in the capacity of the economy to produce goods and services. Growth in per capita real GDP is growth in the economy's output of goods and services per person.

(b) Growth in per capita real output is a measure of growth in the standard of living. Growth in per capita real output depends on both the growth in total real output and the growth in population.

(c) If population is growing, growth in per capita real GDP will be less than growth in potential GDP.

Exercise 25.2

(a) Ten years in the future the country with a growth rate of 3.5% will have potential GDP of \$141 billion, while the country with a growth rate of 3.25% will have potential GDP of \$138 billion, a difference of 2.2 percent.

(b) Twenty years in the future the potential GDPs will be \$200 billion and \$190 billion respectively, a difference of 5.3 percent.

Exercise 25.3

By growth accounting the contributions to annual growth in potential output are:

(a) Labour force growth 1.4%.

(b) Capita stock growth 1.0%.

(c) Improved productivity 1.1%.

Exercise 25.4

By growth accounting the growth in real GDP=0.7(growth in labour force)+0.3(growth in capital stock)=0.7(2.5)+0.3(1.5)=1.75+0.45=2.2%. Capital stock does not grow as fast as labour force with the result that falling labour productivity reduces output per worker.

Exercise 25.5

(a) Annual growth in country A is 2.9% and in country B is 3.9%.

(b) Country A because output growth is greater than labour force growth.

(c) The capital to labour ratio rises in A but falls in B.

(d) Faster growth in total output in B comes from faster growth in the labour force but the fall in the capital to labour ratio in B lowers labour productivity, which is output per worker.

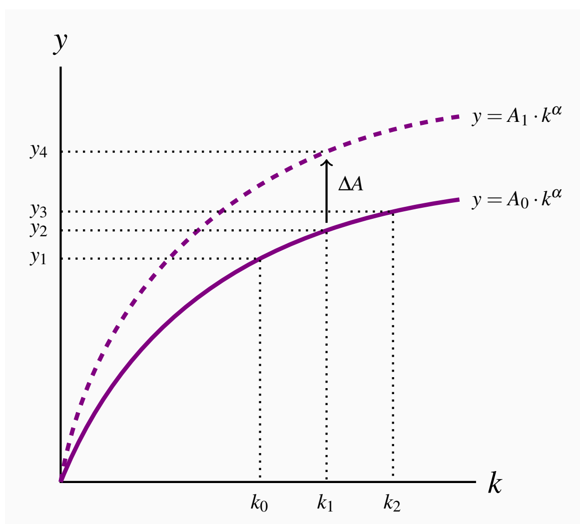
Exercise 25.6

With equal growth rates of capital and labour at 2.5% a year, the capital/labour ratio is constant and the economy enjoys constant returns to scale. Potential output will grow at 2.5% a year.

Per capita output will be constant as output and labour grow at the same rate.

Improved technology that increased total factor productivity by 1.5% a year would result in growth in per capita real GDP by 1.5% a year.

Exercise 25.7



(a) Technology is the key to improved standards of living because increases in output per worker arising from increases in the capital to labour ratio are limited by diminishing returns and eventually fall to zero.

(b) Increasing capital per worker $(K/N = k)$ move the economy along the per worker output function with a decreasing slope caused by diminishing returns to the capital per worker ratio. Increase in k from k_0 to k_2 result in smaller and smaller increase in y.

(c) An improvement in productivity (ΔA) shifts the production function up, raising y at every k and is not subject to diminishing returns. In the diagram, improved productivity increases y from y_2 to y_4 without any change in labour or capital inputs. Productivity growth from new technology has the potential to provide sustained increases in output per worker and standards of living.

www.lyryx.com