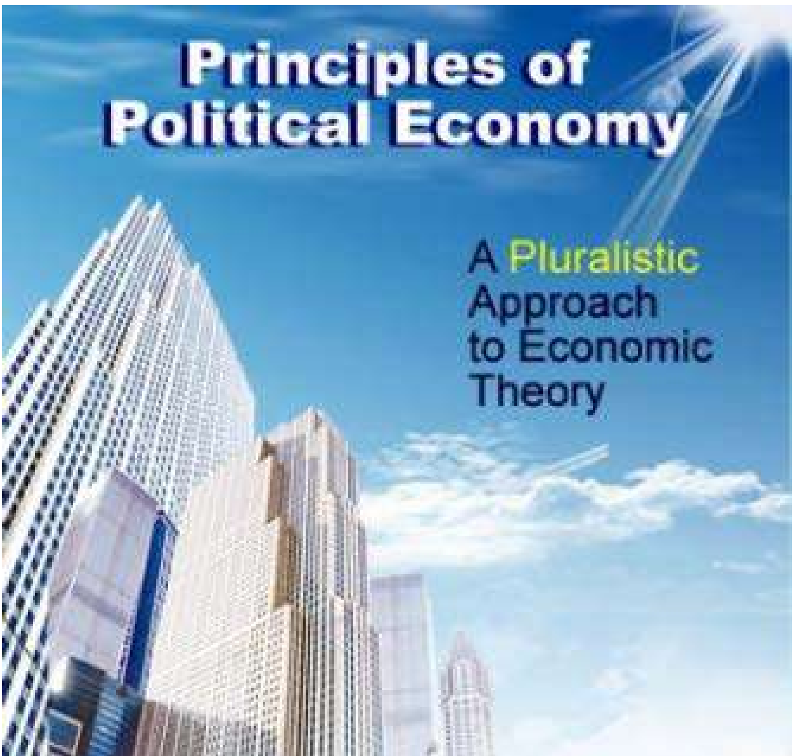
Principles of Political Economy

A Pluralistic Approach to Economic Theory

Daniel E. Saros

2019



Valparaiso University

Valparaiso, Indiana

CONTENTS

PUBLICATION AND LICENSE INFORMATION

Published on June 12, 2019 by Valparaiso University, Valparaiso, IN 46383

Any instructor may adopt this textbook for a course. If you choose to adopt it, please notify the author at Dan.Saros@valpo.edu.

This textbook is an open educational resource (OER). For more information about OERs, please visit: https://www.oercommons.org/

The vision for the cover was developed by Michael J. Saros.

The three images used to produce the cover for the e-book are from pngtree.com:

Vector Designed By Nicolas Perez from

Vector Designed By pngtree from

Vector Designed By アヒルスタジオ from

DEDICATION

FOR MARTIN H. WOLFSON

MY FRIEND AND TEACHER

ABOUT THE AUTHOR

Daniel E. Saros is an Associate Professor of Economics at Valparaiso University, where he has been a member of the faculty since 2004. He received his B.S. in Economics at Bowling Green State University in 1999, and his MA in Economics in 2001 and Ph.D. in Economics in 2004 from the University of Notre Dame. He served as the Director of the International Economics and Finance Master of Science Program from 2012 to 2013 and as the Chair of the Department of Economics from 2014 to 2018 at Valparaiso University. He also served as a member of the Editorial Board of the *Review of Radical Political Economics* from 2013-2016.

Prof. Saros has published two books with Routledge, including *Information Technology and Socialist Construction: The End of Capital and the Transition to Socialism* (2014) and *Labor, Industry, and Regulation during the Progressive Era* (2009). He has also published articles in the *Review of Radical Political Economics* and the *Journal of Economic Education*. His teaching interests include money and banking, the history of economic thought, and economic theory. His primary research specializations include schools of economic thought and methodology, economic education, and economic history.

PREFACE

This textbook is unique among economics textbooks. It contains many of the same topics as mainstream textbooks, but it includes and takes very seriously heterodox critiques and alternatives to the mainstream approach to economics. It includes a whole range of alternative theories, including Post-Keynesian, Austrian, Marxian, radical, feminist, institutionalist, and other approaches. The purpose is to teach students about alternative schools of economic thought but also to deepen their understanding of the dominant, neoclassical approach to economics. In this sense, it draws a great deal of inspiration from Richard Wolff and Stephen Resnick's *Contending Economic Theories*. Following Wolff and Resnick, an even broader objective is to teach students that economics is a discourse and that no single voice can rightfully claim to have a monopoly on the truth about economics. A final objective is to make available this material at no cost to students. Because of the high price of attending American colleges and universities, I have decided to do my small part to make higher education more affordable. Therefore, the e-book is available for free to students. Students who desire a hard copy can use a print-on-demand service to print one.

The approach of this textbook was developed over a period of fifteen years, during which I have been teaching economics at Valparaiso University, and three and a half years prior to that time when I was a graduate teaching assistant at the University of Notre Dame. I owe a great debt to certain professors at Notre Dame, who helped me develop my pedagogical approach, including Prof. Amitava Dutt, Prof. Philip Mirowski, Prof. David Ruccio, and Prof. Martin Wolfson. Instructors at Bowling Green State University, who helped place me on a path towards the teaching of heterodox economics, include Prof. Neil Browne, Prof. Kevin Quinn, and Prof. Steve Ziliak. I had many other excellent instructors at both institutions, but these instructors stand out as major influences with respect to heterodox economics instruction.

Regarding neoclassical economics instruction, Prof. John Hoag is my most important influence. His outstanding intermediate microeconomics course prepared me for so much of what I encountered in graduate school that I owe him a great debt. His impact on me is surely present throughout the microeconomics part of this textbook. I feel the need to mention Prof. Amitava Dutt and Prof. Jaime Ros when I consider how much I learned about mainstream macroeconomics in their graduate classes and as a teaching assistant for Prof. Dutt. Prof. Mary Ellen Benedict also taught me a great deal about econometrics, but this knowledge plays less of a role in this book.

The standard practice as it relates to the writing of economics textbooks is to only cite sources when the knowledge is not regarded as common knowledge in the profession. When concepts and ideas appear in many textbooks without citation, I have not included detailed references. Obvious examples include the production possibilities model and the supply and demand model. Nevertheless, several textbooks have influenced me the most over the years and have influenced my understanding of neoclassical economic models. These textbooks include *Foundations of Economics* by Robin Bade and Michael Parkin, *International Economics* by Robert Carbaugh, *Core Economics* by Eric P. Chiang, *Managerial Economics* by Paul G. Keat, Philip K.Y. Young, and Stephen E. Erfle, *Economics* by Campbell McConnell and Stanley Brue, *The Economics of Money, Banking, & Financial Markets* by Frederic Mishkin, *Principles of Economics* by OpenStax College, and *Economics* by Paul Samuelson and William Nordhaus. These authors taught me a great deal during my time as a student and as an instructor. Even though all the economic concepts presented can be found in multiple textbooks, I have chosen to include more citations of textbook authors than is typically the case to acknowledge the contribution of these authors to my own economics education. When authors are cited, I do not include complete bibliographic information each time I cite them in the book. Instead, I only include the author's name and the year of publication in chapter endnotes throughout the book. A list of references appears at the end of this textbook so that the interested reader can see the full bibliographic information for any of the sources mentioned in the chapter endnotes. This approach greatly limits the amount of information that appears in endnotes and makes the textbook easier to read. Complete information is provided in the endnotes for some sources cited in the book, including news articles, data sources, and references where permissions were required to include the information.

Because I remember where I first encountered specific topics, I have taken care to mention the sources when possible. Frequently the sources are textbooks, but sometimes the sources are lectures that I remember from my days as a student. I have included such references whenever possible.

In terms of mathematics, I have tried to include every step in the solutions to quantitative problems throughout the book so that students need not fill in the blanks themselves.

I have also tried to include all the mainstream concepts and models that typically appear in economics textbooks so that heterodox economics instructors should not feel the need to supplement the textbook with a mainstream textbook.

Daniel E. Saros

Valparaiso, Indiana

March 13, 2019

ACKNOWLEDGMENTS

I began writing this textbook in the summer of 2013 and did not complete it until June 2018. The love and support of my wife Stacy and my son Charles were the most important factors that allowed me to complete the book. Their patience and understanding mean more to me than I can ever express. My parents also provided valuable support and continued to show interest in the project as it progressed. I am deeply grateful to them as well. My brother Michael was exceedingly patient with me as he developed a vision for the e-book cover. I am extremely pleased with the final product, and he deserves all the credit for the vision.

Aside from the many instructors and textbook authors that I mention in the preface to this book, I want to express deep appreciation to Valparaiso University for granting me a University Research Professorship. Freed from teaching responsibilities during the spring of 2018, I managed to complete the five remaining chapters and edit most of the book. Valpo also agreed to publish the textbook as an Open Educational Resource (OER), ensuring that students and interested readers can access the free e-book anywhere in the world where access to the internet exists. I am thankful to Valpo as well for granting me an OER Development Award in April 2019, which allowed me to cover most of the costs that I incurred to obtain permissions for the use of copyrighted material. I am also extremely grateful to Pressbooks for providing the software that allowed me to convert my MS Word files into a format more suitable for publication.

I also wish to thank Liyuan Duan and Yuanyuan Xie, two graduate International Economics and Finance (IEF) students at Valparaiso University who worked diligently for me during the summer of 2013. They edited five chapters and offered valuable suggestions that I incorporated into the final version of the textbook.

I am also grateful to Prof. Jon Bull in the Christopher Center Library at Valparaiso University for helping me decide on the most appropriate license for the book and for other very helpful advice. I also wish to thank Prof. Ruth Connell for some helpful suggestions regarding references and permissions.

Finally, I would like to thank the many students that I have taught over the years. Their reactions to the lessons in this book helped me to refine and improve the presentation in countless ways.

BRIEF CONTENTS

PART THREE: PRINCIPLES OF MACROECONOMIC THEORY

PART FOUR: PRINCIPLES OF INTERNATIONAL ECONOMIC THEORY

COMPLETE CONTENTS

LIST OF FIGURES

Figure 20.20: The Influence of Supply and Demand on Exchange Rates in Marxian Economics

LIST OF TABLES

PART II

PART ONE: AN INTRODUCTION TO ECONOMIC THEORY

CHAPTER 1

THE DISCOURSE OF ECONOMICS

Goals and Objectives:

In this chapter, we will do the following:

1. *Define* the subject of economics
2. *Outline* the history of the discipline of economics
3. *Acknowledge* the dominance of neoclassical economics
4. *Analyze* the method of neoclassical economics
5. *Question* the distinction between positive and normative statements in economics
6. *Investigate* the nature of economic models
7. *Review* the basic mathematical concepts that are required to understand economic models

What is Economics?

It is common for people with very little knowledge of economics to assume that the subject is all about money. Other times people assume that economics is a study of commerce and business relationships. Although money and commerce are topics of great interest to economists, the subject is far broader than these areas. A far more accurate definition of economics identifies the subject as a **discourse**. A discourse is simply a conversation or a discussion. Economics, however, is not just any conversation or discussion. It is a discourse about how human societies function. To be more exact, we may define economics in the following way:

- **Economics** is an intellectual discourse concerned with the manner in which societies produce, circulate, and distribute products and services, as well as the consequences for human welfare that follow.

As the definition indicates, economics is an *intellectual* discourse. That is, it requires the use of the intellect. Economics is about big ideas and requires logical thinking and careful reasoning. Frequently, the conversation takes place within academia. Professors discuss these ideas with students within college and university classrooms. They also discuss these ideas with one another at professional conferences and in academic journals that deal with specialized areas of economics. Economic discussion also occurs among government officials as they debate the most appropriate economic policies at the local, state, and federal levels. For example, public officials discuss whether tax rates should be increased or decreased, or whether to increase or decrease government spending. The discourse of economics, however, is something in which all people participate at some point. The discussion need not be highly sophisticated to be an economic conversation. When a person tells a friend why she thinks the price of an item is higher at one store than at another store, the person and her friend are participating in the discourse of economics.

Economics, therefore, is a subject that explores certain social processes in our society, but it is also an investigation into how social processes affect our well-being. For example, economic discussion might involve questions about how changes in tax policy might influence economic growth or how price differentials might affect consumer welfare. Such questions are central to the subject of economics.

This way of thinking about economics might seem straightforward to the reader, but it turns out that most professional economists, if asked, would offer a very different definition of economics than the discourse definition provided above. How most economists define the subject of economics and the reason this author is proposing the discourse definition of economics will become clear once we briefly explore the history of the discipline of economics.

A Brief History of the Discipline of Economics

It might surprise the reader to learn that economics is a relatively young discipline when considering the entire history of human knowledge. The ancient Greeks studied geometry, moral philosophy, and politics, but they did not have a discipline that they referred to as economics. In fact, it was not until the seventeenth and eighteenth centuries that a distinct discipline emerged that is anything like what we today refer to as economics. At the time, people called the subject "political economy." Economic ideas had been discussed prior to that time, but they were embedded in discussions of politics and moral philosophy, and so no one held a conception of a body of thought referred to as "economics."

It is natural for the reader to wonder why economics emerged at such a relatively late date in human history. The reason is that certain parts of the world were undergoing a massive transformation during those centuries that ultimately helped bring about a revolution in human ways of understanding the social world. This massive shift in human organization coincided with the decline of the **feudal system** that existed in Western Europe prior to that time. The old feudal system had been based on a strict social hierarchy with a king at the top, followed by feudal lords who employed serfs to work on large tracts of land. The king ruled by divine (God-given) right, and each person's place in the social hierarchy was accepted as part of a natural order established by God. Those who questioned it were dealt with harshly, and the system continued in this manner for hundreds of years.

With the growth of markets and long-distance trade, however, the old feudal system began to weaken. As market towns emerged, the surplus income of feudal lords was increasingly spent on items that could only be obtained by merchants engaged in long distance trade. As individual property rights began to spread and commerce flourished, those with the means to do so began to assert property rights over the land. The feudal serfs, who once worked the land in accordance with religious custom and tradition, were now being driven from the land. The so-called enclosure movement, which first began in England and then spread to other parts of Western Europe, involved the forceful expulsion of serfs from privately owned plots of land as private property in agriculture began to spread. Forced from the land, the former serfs made their way to the towns, which grew into large industrial cities. There they sought work in the factories and the mines as wage laborers. Their search for paid work reinforced the spread of private ownership of the means of production. In this manner, **market capitalism** (a system based on the private ownership of land and the means of production as well as the voluntary exchange of products) gradually displaced the old feudal system.

With the growth of markets and long-distance trade, most people's livelihoods came to depend for the first time in human history upon movements in prices and wages. It is difficult to overstate the magnitude of this change. Never before had the majority of people experienced anything quite like it. Now they were at the mercy of movements in monetary variables that appeared to follow no clear logic of their own. As unpredictable as the weather, people struggled to understand why and how these variables moved as they did. As a result, they were becoming conscious of something called "the economy," which was a subject that prior to that time simply did not exist. That is, the act of producing, acquiring, and consuming previously had been deeply embedded in other aspects of social life. These acts had been a part of the religious order and traditional ways of living. Now they appeared to have a separate life of their own, independent of other areas of life, whether religious, cultural, or familial. People struggled to understand the changes. Scholars emerged, claiming to have answers. These scholars were the first political economists.

For a fuller discussion of the first political economists, the reader should consult a history of economic thought textbook. Certain key figures should be mentioned at this stage, however, to justify the discourse definition of economics provided in the previous section. Without question, the person that most professional economists today regard as the father of modern economics is the Scottish political economist and philosopher Adam Smith (1723-1790). Smith's most famous work, published in 1776, is titled, *An Inquiry into the Nature and Causes of the Wealth of Nations*. The metaphor for which Smith is famous is introduced in this book and is commonly referred to as the **invisible hand** of the market. According to this notion, within an economic system that is based on private property and voluntary market exchange, each person is led as if by an invisible hand to pursue the social interest. That is, even though everyone pursues only their own self-interest and no one intends to serve the social interest, the social interest is served nevertheless. To understand the reason, simply consider a person who wishes to receive a haircut. If the person pays a hair stylist $10 for a haircut, then the person obtains the desired haircut and fulfills his or her self-interest. At the same time, the reader should notice that the hair stylist's primary aim is to obtain an income. When the customer pays the hair stylist the $10, the customer serves the interest of another person (i.e., the hair stylist) even though it was not his or her intention to do so. Similarly, by pursuing an income, the hair stylist serves the interest of the customer by providing the haircut even though serving the interest of the customer was not the primary aim of the hair stylist. Both the buyer and seller of the haircut serve each other's interest in this case even though each only pursues his or her own interest. To use Adam Smith's phrase, each market participant is led as if by an invisible hand to serve another person.

Another major figure in the history of political economy is the English political economist and financier David Ricardo (1772-1823). Ricardo's most famous work, published in 1817, is titled, *The Principles of Political Economy and Taxation*. Ricardo developed many very influential ideas within this relatively small treatise. One of the most important theories Ricardo developed is today the bedrock of international trade theory. This theory, known as the theory of comparative advantage, holds that countries may benefit from international trade even if one of the countries is better able to produce both the product that it is selling and the product that it is buying from the other country. This startling result suggests that a rich and powerful nation may benefit from trade with a relatively weak nation, and vice versa. Mutual gains from trade thus extend to the international marketplace, according to Ricardo. In Chapter 19, we will consider Ricardo's argument in detail and demonstrate the conditions under which his claim holds true.

Although Smith and Ricardo offered generally positive assessments of the new market capitalist system that had developed in Western Europe, not all political economists shared their optimism. One final political economist that we will consider is the German political economist and revolutionary socialist Karl Marx (1818-1883). Marx is famous for several works, including *The Communist Manifesto* (1848) and *Capital* (1867-1894). *Capital* (which is written as *Das Kapital* in German) is a multi-volume work that contains Marx's most sophisticated theory of market capitalism. According to Marx, the new capitalist system was leading to the most rapid development of human productive technology that the world had ever known. At the same time, these developments were occurring within the context of an economic system in which a majority of the population (the working class or proletariat) was being exploited by a small minority of property owners (the capitalist class or bourgeoisie). In Marx's view, the struggle between these social classes was the source of economic crises and many other social problems characteristic of the new economic system.

In addition to these political economists, many others wrote long treatises describing in detail their conclusions about the advantages or disadvantages of the new economic system. Around these major thinkers, schools of economic thought developed that aimed to refine and extend the ideas of each major thinker. It will now be easier for the reader to understand why the discourse definition of economics offered in the previous section is the most appropriate one. The clashes between economic ideas and the depth of the disagreements between different schools of economic thought were so great that they were not easily resolved. Conversations and discussions often have participants with very different viewpoints and when the discussion is about something as complex as the functioning of human societies, one cannot expect to find easy resolutions. Therefore, by referring to economics as a discourse, we are being honest about the challenges that face human understanding and how those challenges produce different and competing conceptions of the world in which we live.

The Dominance of Neoclassical Economics

Given what has been said about the history of political economy and the reasons for adopting a discourse definition of economics, the reader might expect economics textbooks to be filled with many different economic worldviews and opinions as represented by different schools of economic thought. In fact, the opposite is true. Readers of introductory economics textbooks are introduced to the ideas of a single school of economic thought. Towards the end of the nineteenth century, this school of economic thought became the dominant voice in the economics discourse, particularly within the western capitalist nations. In the 1870s, three political economists working independently in France, Britain, and Austria developed ideas that would form the foundation of a school of economic thought that shared many features with the classical political economy of Smith and Ricardo. At the same time, the methods and concepts used were sufficiently new and different to warrant the new title of **neoclassical economics**. By the end of World War II, this school of economic thought had become so dominant in the United States that today students are introduced to economics as if it is a single school of economic thought without any challengers at all. During the Cold War, Marxian economics continued to be the dominant school of economic thought in the command socialist economies such as the Soviet Union. Now that the Cold War is over, however, neoclassical economics reigns supreme. Its dominance is the reason that it is sometimes referred to as **mainstream economics** or **orthodox economics**.

How then should we define neoclassical economics? The following definition places emphasis on the dominance of neoclassical economics even as it includes the key elements of the definition of economics that most professional economists accept.

- Neoclassical economics is the dominant school

> of economic thought that defines economics as the social science concerned with the efficient use of scarce resources to achieve the maximum satisfaction of unlimited individual wants.

Neoclassical economists, therefore, define economics in a manner that is much narrower than the discourse definition provided earlier in this chapter. Neoclassical economists define economics first and foremost as a science, much like physics, and in fact many professional economists have welcomed the analogy. Indeed, many of the concepts that ultimately became part of neoclassical economics were transferred to economics directly from physics.[1] The definition of economics as a science also indicates that it is a subject in which only experts can participate in a meaningful way. It, therefore, restricts the growth of participation in the discourse of economics insofar as many people decide that they simply do not have the expertise to form their own opinions about the subject or to challenge experts on economic issues.

In addition, it should be clear that neoclassical economists strictly limit the scope of their inquiry with their definition of economics. The focus on efficiency in the use of scarce or limited resources is far more specific than the focus on the social processes of production, circulation, and distribution that are emphasized in the discourse definition of economics. Should efficiency (or obtaining the most from the least) really be the primary focus of economics? Should equality also be a concern of economists? If not, why not? Furthermore, the neoclassical definition of economics makes a major claim about human beings that is absent in the discourse definition. It is asserted that humans have wants that are unlimited. Is there really no limit to human desires? Will the answer be the same for all people, at all times, and in all places? These aspects of the neoclassical definition raise serious questions as to whether efficiency should be the exclusive focus of economics and whether human wants really are unlimited.

Regardless of our answers to these questions, it is important to understand that neoclassical economists dominate modern discourse about economic issues. Furthermore, that dominance has become so extensive that introductory textbooks almost universally refuse to acknowledge that the perspective represented corresponds to a single school of economic thought. The great majority of economics textbooks have titles that refer to economics only and almost never to neoclassical economics. Moreover, the textbooks include virtually no discussions of the history of the discipline so that students may consider how the ideas were developed through a process of historical debate and discussion. Students do not learn, for example, that many other approaches to economics exist, including Marxian economics, Austrian economics, Post-Keynesian economics, Sraffian economics, feminist economics, and institutionalist economics. This very diverse group of schools of economic thought that reject at least some part of the neoclassical approach to economics may collectively be referred to as **heterodox economics**. Because this textbook is subtitled, *A Pluralistic Approach to Economic Theory*, the reader may look forward to lengthy discussions of many different approaches to economics.

These different schools of economic thought frequently have little in common except for their rejection of neoclassical economics. It is that opposition to the dominant school of economic thought that really binds them. Furthermore, the reader should not assume that the distinction between neoclassical economics and heterodox economics is one of right versus left in the political sense. While neoclassical economics is firmly on the right side of the political spectrum and Marxian economics is firmly on the left side, Austrian economics is even further right than neoclassical economics. It is helpful to understand the political orientation of each school of economic thought, but one should not assume that a rejection of neoclassical economics is driven only by different perspectives on the role of government in the economy. Frequently, methodological differences are behind a school's rejection of the neoclassical approach to economics, which is a subject to which we now turn.

The Entry Point and Logic of Neoclassical Economics

As explained in the previous section, today most professional economists are members of the neoclassical school of economics. For that reason and because heterodox schools of economic thought frequently define themselves in relation to the dominant school of thought, we give neoclassical economics extensive treatment in this chapter and throughout the book. In this section, we consider the method that most professional economists use.

If asked about the method that they use to acquire knowledge in their discipline, most economists will quickly mention the **scientific method**. As every high school science student knows, the scientific method has several key steps: the formulation of a question, the formation of a hypothesis, the making of predictions, the testing of the predictions against observed reality, and the analysis of the results. If we were operating with the neoclassical definition of economics, which defines economics as a science, then we would certainly identify the scientific method as the method of economics. Because we have adopted the discourse definition of economics, however, we will describe the method of economics rather differently. Our objective will be to identify the method of neoclassical economics in a manner that is compatible with this definition.[2]

Whenever a person enters a discourse with another person, she brings to the discussion certain bedrock assumptions. These foundational concepts are sometimes mentioned explicitly, but frequently they are introduced without the other participants in the discourse, and perhaps not even the person introducing them, being fully aware of them. For example, a person who argues that tax rates ought to be reduced may implicitly assume that people have a moral claim to any income they earn in the marketplace. That people have moral rights of this kind may be an implicit assumption that the person holds that he has never critically examined. Nevertheless, it will be used as the foundation upon which many arguments will be constructed as that individual participates in the discourse in which he is engaged. To operate without such assumptions is like trying to make an argument without any starting point at all. Can you imagine trying to proceed in this fashion?

Economists also possess certain bedrock assumptions whenever they are engaged in the discourse of economics. When discussing the bedrock assumptions of a school of economic thought, we will refer to those assumptions as the **entry point**. An entry point serves as the starting point in a discourse. Each school of economic thought has an entry point. For neoclassical economists, the entry point is physical and human nature.[3] That is, neoclassical economists make certain assumptions about the physical world and about the nature of human beings that serve as the foundation upon which they erect a complex economic worldview.

The point of entry of neoclassical economics can be broken down into three parts as follows:[4]

1. Society's endowment of resources is taken as given.
2. Society's production technology is taken as given.
3. The individual preferences of every person in society are taken as given.

The first two components of the neoclassical entry point relate to the physical world and so we will discuss these together. Neoclassical economists assume that our society is endowed with a specific amount of resources (i.e., a resource endowment) that can be used to produce products and services. They also assume that our society has a specific amount of knowledge of the methods that may be used to transform those resources into products and services (i.e., technology).

Both the resource endowment and the production technology are taken as given. What does it mean to state that these elements of the physical world are taken as given? It means that the specific amount of available resources and the available production technology are the bedrock assumptions that will form the starting point for any further analysis. Furthermore, neoclassical economists do not aim to explain how society came into possession of this combination of resources or this knowledge of how to produce products and services. It is here that the disciplinary boundaries are sharply drawn so that only certain questions are to be regarded as *economic* questions. For example, a neoclassical economist would not consider a question about how the United States acquired its territory to be an economic question. The neoclassical economist would consider this question to be one for historians to answer but not economists. Whether the territory of the United States was acquired through conquest or through purchase is not something that a neoclassical economist would consider important in terms of answering strictly economic questions. Similarly, a neoclassical economist would not consider a question about how we came into possession of modern information technology to be an economic question. It would also be regarded as a question for historians but not for economists.

Similarly, the third component of the neoclassical entry point takes individual human preferences as given, which is an assumption about human nature. That is, each person is assumed to have certain preferences for products and services, and the origin of these preferences is beyond the scope of neoclassical economics. For example, a question about how a smoker developed a preference for smoking cigarettes would be regarded as a non-economic question within the neoclassical economics profession. This question might be one for psychologists or marketing experts to explore, but economists simply accept that people have preferences and then build their arguments using that starting point. As the reader can observe, the neoclassical entry point places strict limits on what can be regarded as falling within the purview of economic discourse.

To successfully participate in a discourse, it is necessary also to possess a **logic** or method of reasoning.[5] Successful argument requires not only the building blocks of analysis but also the ability to link them together in a way that creates a unified structure.[6] The logic of neoclassical economics may be referred as one of **unidirectional causality** (or cause-and-effect).[7] According to this logic, one variable always affects another variable in a unidirectional manner. That is, variable A causes a change in variable B, but it is never the case that variable B causes a change in variable A. The direction of causality only runs in a single direction. This method of proceeding lends itself well to mathematical reasoning, as we shall observe later in this chapter. Nevertheless, we can imagine situations in which this logic may be restrictive. For example, consider the claim that consumer preferences directly affect the price of diamond rings. According to this simple theory, as consumers desire more diamond rings, the price is driven upwards, other factors held constant. Is it possible, however, that the causal connection might be reversed at times? That is, could a rise in the price of diamond rings cause some consumers to desire more of these items? It may be possible if consumers rely on price as an indicator of the quality of the item. If so, then prices may influence preferences just as preferences influence prices. This situation involving **mutual causality** between variables[8] is not consistent with the neoclassical logic and such cases are strictly forbidden within neoclassical economic theory. As we will observe later, practitioners of other schools of economic thought have sometimes applied this type of logic in the construction of their theories.

The Positive/Normative Distinction in Neoclassical Economics

Another major aspect of the neoclassical method relates to the distinction between positive statements and normative statements in economics. A **positive statement** is supposedly purely factual and without any value or moral content. For example, the statement, "Dave is an American citizen," would be considered a descriptive statement, which may be either true or false. A **normative statement**, on the other hand, is a value-laden statement that is neither true nor false, regardless of how interesting and important to an individual it may be to determine whether one agrees or disagrees with the statement. For example, the statement, "U.S. officials should pursue a policy of full employment," would be considered a normative statement.

The distinction has its roots in the ideas of the famous Scottish philosopher and historian David Hume (1711-1776) who famously wrote in his *Treatise of Human Nature* (1738) that "No 'is' implies an 'ought.'" According to Hume, no matter how much we might investigate a fact, we can never extract any moral content from it. The problem, of course, is that if we cannot derive moral implications from facts then how can we ever acquire any moral knowledge at all? This manner of thinking has become central to the neoclassical method of reasoning. According to neoclassical economists, because moral knowledge is impossible, the only purely scientific statements are positive statements. Neoclassical economists aim to generate meaningful positive statements in economics (using their unique entry point and logic) and then determine whether they are true or false. They are also interested in normative questions, but they do not consider those questions to be ones that can be answered once and for all. Similarly, they do not consider normative statements to have correct or incorrect answers. Of course, if we agree on certain ends (e.g., full employment *should* be our goal), then positive economics can help us reach that goal. If we disagree about the ends, however, we will never resolve our normative debate.

Many people learn at a young age about the difference between fact and opinion. This distinction might seem to be an easy one for the reader to accept. A critical examination of the distinction between these two kinds of statements, however, reveals that it is more problematic than one might initially believe. For example, consider the following statement: "John is not unemployed." At first glance, this statement might appear to be a purely positive statement. Let's add a bit of context to the statement, however, and then reevaluate it. Consider the following facts surrounding the statement:

1. John lost his job about 13 weeks ago.
2. After losing his job, John searched for a new one for about 7 weeks with great effort.
3. About 6 weeks ago, John became so frustrated with his failed job search that he gave up looking for a job. He now sits at home and watches TV all day as his bills pile up.

Returning to our earlier statement that John is not unemployed, the reader might be surprised to discover that the statement is true. The reason is that the U.S.

Bureau of Labor Statistics (BLS), which publishes the official unemployment rate each month for the U.S. Department of Labor, defines unemployment such that a person must have been actively searching for a job within the last four weeks to be considered unemployed. A worker like John, who has given up on his search, is officially counted as being outside of the labor force rather than unemployed. According to the BLS definition of unemployment, John is not unemployed.

It should be clear now just how much normative content the statement really possesses. Many people might look at this example and argue that John should be counted as unemployed. After all, he wants a job and cannot find one. Other people will look at this situation and argue that we must exclude John because he has given up his search. From a practical standpoint, how are we to count frustrated workers like John when they are not even making a token effort to find work? The point here is not to argue that John should be counted as employed or unemployed. Instead, the point is that a statement that appears, on the surface, to be entirely descriptive, in fact contains a great deal of normative content.

The reader might wonder whether this example is a far-fetched one that is unlikely to arise. On the contrary, examples such as these are not difficult to find within the discourse of economics. Many economic variables are constructed by economists, and in each case, the theorist must decide which elements to include and which elements to exclude. The very act of selection is a value-laden decision. Once constructed, any statements that refer to changes in these variables may appear to be unambiguously positive statements when in fact they are anything but purely descriptive.

For this reason, it is important to be cautious regarding claims to objectivity in economics. We close this section with words from a famous economist by the name of Joan Robinson who had a habit of challenging many aspects of mainstream economics. In Robinson's words,

> There has been a good deal of confused controversy about the question of 'value judgments' in the social sciences. Every human being has ideological, moral and political views. To pretend to have none and to be *purely objective* must necessarily be either self-deception or a device to deceive others. A candid writer will make his preconceptions clear and allow the reader to discount them if he does not accept them. This concerns the professional honour of the scientist.[9]

The Ceteris Paribus Assumption in Neoclassical Economic Models

Armed with the neoclassical entry point and a unidirectional causal logic, neoclassical economists develop theories to explain different aspects of economic life. These theories are expressed with the use of formal **economic models**. Economic models are quite like physical models in that the purpose of each type of model is to simplify a complex reality. For example, a model airplane is designed to capture the key features of an actual airplane. To serve as a good model airplane, the model should include a propeller, the landing gear, the fuselage, a rudder, wings, a windshield, and other component parts. If any of the essential components are missing, then the model is not a good model. On the other hand, the model should exclude many inessential elements that actual airplanes possess. In the extreme case, if a model included every element of an actual airplane, then it would cease to be a model and would instead be an actual airplane. As a result, the model builder must make some careful choices about which elements to include and which elements to exclude from the model.

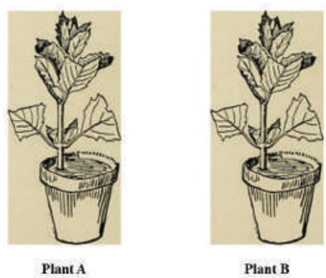
Economic models are very similar to physical models in that both represent simplifications of complex realities. Economic models, however, are abstract in nature, which means that they generally do not possess a physical form. Instead, they are mental constructs that theorists design. Aside from that important difference, they are much like the physical models with which the reader is no doubt already familiar.

One point that the reader should consider at this stage is how likely moral judgments are to creep into neoclassical analysis. All theorists confront an infinite amount of data when they begin to construct theories to explain various phenomena. Because the theorist must choose which elements to include and which elements to exclude from the model, it is inevitable that the personal values of the theorist will influence these choices.[10] Once again, we see that even theories that are framed in purely positive or descriptive terms will inevitably contain implicit moral content. It is the job of the astute observer to detect this normative content when it is not explicitly acknowledged by the model-builder.

Before we look more closely at the construction of neoclassical economic models, it is essential to introduce the key assumption that is present in all such models. The assumption that is often only implicit in neoclassical models is the **ceteris paribus assumption**. This assumption, also referred to as the *other-things-equal assumption*, allows a theorist to hold all other variables constant so that she can focus only on the relationship between the variables in which she is most interested.

An example from the natural sciences will help clarify the critical role of the ceteris paribus assumption in the construction of neoclassical economic models. For example, consider a biologist who performs an experiment using two different but identical plants, as shown below in Figure 1.1.

Figure 1.1: Two Identical Plants



Source: Oliver, George Watson and Alfred Carl Hottes. Image from *Plant culture; a working handbook of every day practice for all who grow flowering and ornamental plants in the garden and greenhouse.* A.T. De La Mare co., Inc. New York, 1921. p. 17. www.flickr.com

The biologist's purpose is to test the hypothesis that a moderate amount of sunlight encourages plant growth. Suppose that during the month, the biologist gives each plant the same amount of water and the same amount of plant food. Assume, however, that the biologist gives Plant A a moderate amount of sunlight and Plant B zero sunlight. We might expect Plant A to grow and thrive whereas Plant B dies, thus providing evidence in support of the hypothesis. The conditions chosen for this experiment as well as the observed result are represented in Table 1.1.

Table 1.1: The Conditions and Results for Experiment 1

	Plant A	Plant B
Conditions	Some plant food	Some plant food
	Some water	Some water
	Some sunlight	Zero sunlight
Observed Outcome	Plant A grows	Plant B dies
Conclusion	Moderate sunlight encourages plant growth.	

Now suppose that the experiment is conducted once again with two new plants that are identical to the ones used in the previous experiment. This time, however, the biologist gives each plant the same amount of plant food, but Plant A is given zero water whereas Plant B is given a moderate amount of water. As before, Plant A is given a moderate amount of sunlight and Plant B is given zero sunlight. The conditions chosen for this experiment as well as the observed result are represented in Table 1.2.

Table 1.2: The Conditions and Results for Experiment 2

	Plant A	Plant B
Conditions	Some plant food	Some plant food
	Zero water	Some water
	Some sunlight	Zero sunlight
Observed Outcome	Plant A dies	Plant B dies
Conclusion	Moderate sunlight does not encourage plant growth.	

The reader might notice that the reason the observed outcome fails to support the hypothesis in the case of Experiment 2 is that other conditions differ for the two plants other than the variable in which the biologist is most interested. That is, because Plant A received zero water, it no longer matters that it received a moderate amount of sunlight. In other words, the hypothesis can only be tested when all other conditions are the same across the two plants except for the amount of sunlight. Without all other conditions being the same, we lose our ability to draw conclusions about the effect on plant growth of the variable in which we are most interested because these other conditions may interfere with plant growth as well.

This manner of proceeding appears to provide neoclassical economists with a way to test economic hypotheses. For example, suppose a neoclassical economist wishes to test the hypothesis that tax cuts in the United States stimulate economic growth. All the economist needs to do is find another economy just like the United States and then convince the government in one country to cut taxes by the prescribed amount while convincing the government in the other country to keep taxes at a constant level. It is also necessary that nothing else change in the two economies during the period of this investigation. If economic growth occurs in the country where the tax cut occurred, then the hypothesis has been supported by the evidence. The reader will, without a doubt, recognize the absurdity of this example. No economy in the world is identical to the United States economy, and even if one existed, it would not be possible to alter only the variables of interest while holding all other variables constant. The biologist, therefore, has a great advantage over the economist in that the biologist may perform a *controlled experiment*.

That is, the biologist can impose the conditions that she finds most suitable to the testing of her hypothesis. Economists generally cannot perform these kinds of experiments.

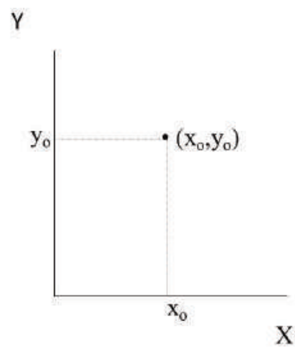
To solve the problem, the economist must rely primarily on *thought experiments*. That is, the economist will imagine that all other variables are held constant except for those that are of greatest interest to the economist. Then allowing the variables of interest to change, she will consider what happens to other variables of interest in the model. Historical economic data and statistical tests may then be used to test economic hypotheses. These statistical tests allow the economist to hold other variables constant to check for the influence of one variable upon another. Clearly, however, these types of tests provide much less convincing evidence than the type that is acquired in a laboratory setting where conditions may be monitored and controlled directly. To the extent that neoclassical economists wish to have their science compared favorably to other sciences, such as physics, it should be clear that the case is a difficult one to make.

The Centrality of Graphical Analysis in Neoclassical Economics

As we have learned, the logic of neoclassical economics requires reference to causal relationships between economic variables. Graphical analysis lends itself well to discussions of relationships between variables, and so the two-dimensional Cartesian coordinate system is widely used in neoclassical economic theory.

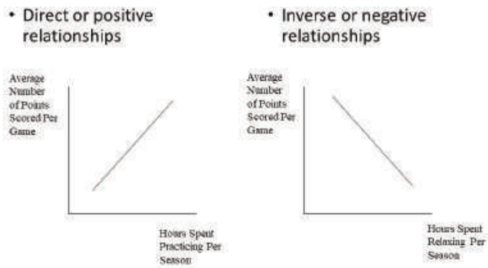
The reader might recall that a two-dimensional graph makes it possible to create an image of the way that two variables relate to one another. In addition, an ordered pair may be used to represent any point in a two-dimensional space. With an ordered pair, the coordinate on the horizontal axis is always listed first and the coordinate on the vertical axis is always listed second, as shown in Figure 1.2.

Figure 1.2: Representing Points in a Two-Dimensional Space



Relationships between variables may also be represented with lines connecting a series of ordered pairs in this same space. When both variables move in the same direction along a line, it is said that they are **positively related variables** or **directly related variables**. When one variable increases as another variable declines, it is said that they are **negatively related variables** or **inversely related variables**. Figure 1.3 provides examples of positively related variables and negatively related variables.

Figure 1.3: Relationships between Variables



Suppose that the ordered pairs on the line in the graph depicting a positive relationship represent combinations of the average points scored by basketball players per game and the hours they spend practicing per season. What is the causal relationship between these variables, according to the graph? The reader might conclude that an increase in hours spent practicing per season causes an increase in the average points scored per game. A moment's reflection reveals, however, that the answer may not be so straightforward. For example, isn't it possible that players score more points due to greater natural ability and that these players practice more because people generally like activities at which they excel? In that case, the higher average number of points scored per game causes players to practice more and the causal relationship is reversed. The point is that we cannot observe the causal connection between these variables. All we observe when we look at the graph is a correlation between the variables.

A similar question might be raised about the inverse relationship depicted in Figure 1.3. The reader might conclude that the cause of lower average numbers of points scored per game is that the players are spending more hours per season relaxing. It is possible, however, that players with lower average numbers of points scored per game due to less natural ability decide that basketball is not very enjoyable anyway and decide to relax more. In fact, both causal arguments may be true to some degree. The point is that causal relationships are not directly observable and so we only observe correlations between variables.

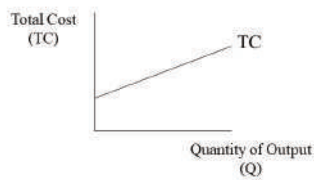
Once again, this insight stems from the work of the philosopher David Hume. Hume argued that we can never observe a causal connection. For example, suppose we are looking at the balls on a pool table, and we observe one ball strike another. After the first ball contacts the second ball, we observe the second ball moving away from the first ball. We repeat the experiment and observe the same result. If we repeat the experiment hundreds, thousands, even millions of times, we always observe the same result. We conclude that the one ball *causes* the other ball to move. Hume's great insight is that we never actually observe this causal connection, only a correlation. No matter how many times we witness the second ball moving after the first ball strikes it, we can never know that the next time the first ball strikes the second ball that the second ball will move. This insight is now known as Hume's critique of induction. Furthermore, any claim that the second ball moves because the first ball strikes it is a conclusion that only the observer (or the theorist) can draw. It is the theorist who imposes causal connections upon observed correlations. Those causal connections are never observed but always imposed. As all students of science should know, *correlation does not imply causation*.

A Neoclassical Model of Production Cost

When neoclassical economists construct theoretical models, they assert causal relationships between variables in accordance with the neoclassical logic of unidirectional causality. The variables that are regarded as the causal variables are referred to as **independent variables**. Similarly, the variables that are regarded as effects in neoclassical models are referred to as **dependent variables**. In other words, the dependent variables *depend upon* the independent variables.

As an example, consider the relationship between the total cost (TC) of production and the quantity (Q) of output. In neoclassical economic theory, it is claimed that total cost depends on the level of output and so total cost is treated as the dependent variable and output is treated as the independent variable. The claim can be stated more precisely as TC = f(Q). According to this mathematical statement of the relationship, total cost is a function of output. This function, which plays a key role in neoclassical microeconomic theory, is known as the **cost function**. Neoclassical economists assert that the relationship is a positive one for a single firm. That is, as production increases, the total cost of production rises. Figure 1.4 represents the relationship as both linear and positive.

Figure 1.4: A Graph of the Total Cost (TC) Function

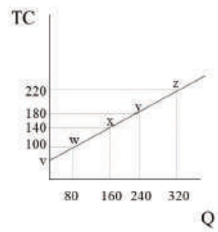


Frequently, students have a difficult time placing variables other than x and y on the horizontal and vertical axes. Because we are interested in economic applications, we will generally use variables other than the highly abstract x and y variables. That is, the variables on the axes measure specific economic quantities. In this case, we are measuring the daily cost of production in dollar terms and the daily production of output in physical terms (e.g., boxes of cereal).

In Figure 1.5, a series of points may be found in the graph with a straight line connecting the points.

Figure 1.5: The Relationship between Output and Total Cost

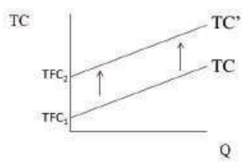
Q	TC	Point
0	60	v
80	100	w
160	140	x
240	180	y
320	220	z



Each point represents a combination of daily cost and daily output. Point v is especially significant. At point v, daily output is zero and yet the firm still incurs a positive amount of cost. That is, the firm incurs $60 of cost even though its output level is zero. How is this combination possible? Well, firms must incur costs even when they do not produce any output. For example, a firm that has a factory must still pay rent. Because this amount of cost does not appear to depend in any way on the amount of output produced, it is referred to as **total fixed cost (TFC)**. As the level of output rises, total production cost also rises and so any amount over and above the fixed cost of production is referred to as **total variable cost (TVC)**.

It should also be noted that this claim that a positive relationship exists between total cost and output assumes that other variables are held constant. For example, a change in rent would certainly affect total cost even though the output level remains unchanged. A rise in rent, for example, would increase the cost of fixed capital assets for the firm and would increase total cost at every output level. The cost curve would thus shift in an upward direction. The ceteris paribus assumption would be violated in that case, as shown in Figure 1.6, and the upward shift would be the result of a change in the firm's total fixed cost.

Figure 1.6: An Upward Shift of the Total Cost Function



It is also possible to write the cost function that exactly fits the data shown in Figure 1.5. To accomplish this task, it is necessary to first determine the slope of the cost function, which can be achieved using any two combinations of output and cost from the table in Figure 1.5. For example, if we use points x and y, then we can calculate the slope as follows:

$$\frac{\Delta TC}{\Delta Q} = \frac{180-140}{240-160} = \frac{40}{80} = \frac{1}{2}$$

The slope in this case is 1/2 or 0.50, which means that a 1 unit increase in output is always associated with a \$0.50 increase in total cost. Because the slope tells us the additional cost that the firm incurs when production increases by 1 unit, it is referred to as the **marginal cost** of production. As explained in the chapter on production technology and cost, a constant marginal cost at every level of output is not to be expected, and neoclassical economists typically make a different assumption about the impact of output changes on marginal cost. We will return to this topic in Chapter 7.

Next, we write the cost function using the following form:

$$TC = TFC + \frac{\Delta TC}{\Delta Q}Q$$

In this function, TFC represents total fixed cost and (ΔTC/ΔQ)Q represents total variable cost. TFC is also the vertical intercept in the graph of the cost function. As mentioned previously, the slope is ΔTC/ΔQ. Finally, TC and Q remain as variables in the function. To determine the amount of total fixed cost, we can pull that information directly from the table in Figure 1.5. That is, when the output level is zero, the cost of production is $60. If that information was not directly available in the table, then that information could be obtained by plugging the slope into the cost function and using any single point in the table. For example, if we use point w, then we obtain the following result:

$$100 = TFC + \frac{1}{2}(80) \Rightarrow TFC = 60$$

With this information, we can now write the cost function that fits the data as follows:

$$TC = 60 + \frac{1}{2}Q$$

Additionally, it is possible to use this cost function to carry out **economic forecasts** of expected future levels of production cost. That is, it is possible to predict the level of cost for other levels of output, assuming this linear relationship continues to hold. For example, if we want to predict the cost of production when output is 120, then we can simply plug the output level into the cost function that we just found as follows:

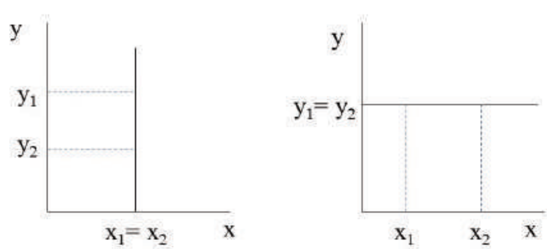
$$TC = 60 + \frac{1}{2}(120) = 120$$

That is, the predicted level of cost is $120. In this example, the total fixed cost and the total variable cost are both $60.

A Review of Basic Mathematical Concepts

In this section, a few basic mathematical concepts will be reviewed. We will make use of these concepts in later chapters and so it is important to have a firm grasp of them. In this chapter, we have discussed the concept of the slope of a straight line. At times, we will encounter straight lines that are either perfectly horizontal or perfectly vertical as shown in Figure 1.7.

Figure 1.7: Two Extreme Cases



The reader might recall that the vertical line has a slope that is undefined (or one that we might say is infinite). Similarly, the reader might also recall that the horizontal line has a slope that is equal to zero. It is important, however, to know also why these slopes are the correct ones. To prove that these slopes are the correct ones, we need only select two points on each line and calculate the slope. For example, if we select points (x_1, y_1) and (x_2, y_2) on the vertical line and then calculate the slope, we obtain the following:

$\frac{\Delta y}{\Delta x} = \frac{y_2 - y_1}{x_2 - x_1} = \frac{y_2 - y_1}{0} \approx \infty$ (or undefined)

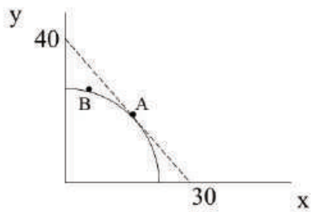
Similarly, if we select points (x_1, y_1) and (x_2, y_2) on the horizontal line and then calculate the slope, we obtain the following:

$\frac{\Delta y}{\Delta x} = \frac{y_2 - y_1}{x_2 - x_1} = \frac{0}{x_2 - x_1} = 0$

At other times, we will encounter lines that are not straight. These nonlinear curves do not have constant slopes but rather variable slopes. That is, the slope varies all along the curve in question. Readers who have taken a differential calculus course know that these slopes can be calculated using differentiation techniques.

In this book, we use a simpler approach to measuring slope at a point on a nonlinear curve. According to **the tangent line method**, the slope of a nonlinear curve at a specific point is equal to the slope of the *unique* straight tangent line that just touches the curve at that point. For example, Figure 1.8 shows an example of a nonlinear curve.

Figure 1.8: The Slope of a Nonlinear Curve

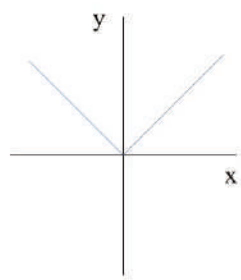


In Figure 1.8, a unique straight tangent line is drawn through point A such that it just touches the nonlinear curve at that point. The slope of the curve at point A is easily computed by calculating the slope of the tangent line as follows:
$\frac{\Delta y}{\Delta x} = \frac{-40}{30} = \frac{-4}{3}$

The reason that the straight tangent line must be unique is that without uniqueness the slope may be impossible to calculate. An example illustrates the point. Consider the absolute value function, which has the following form:
$y = |x|$

Figure 1.9 shows how the graph of this function appears in a two-dimensional space.

Figure 1.9: A Graph of the Absolute Value Function

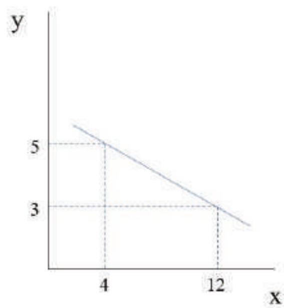


If the reader tries to draw a straight tangent line that just touches the curve at point (0,0), also known as the origin, to determine the slope, then it should become clear that this method fails to provide a satisfactory result. It is true that one can draw a straight tangent line that just touches the curve at the origin. The problem, however, is that many different tangent lines (in fact, an infinite number) can be drawn that just touch the curve at the origin. Therefore, a *unique* straight tangent line cannot be drawn that just touches the curve at the origin. For this reason, the slope is not defined at that point. In the language of calculus, the curve is not differentiable at the origin. At different points throughout this book, a curve will be drawn that contains a sharp corner or edge like the one in Figure 1.9. When such examples arise, the reader should recall that the slope is not defined at these points.

Finally, it will sometimes be necessary to determine the equation of a straight line when only given two points.

For example, suppose you are given the points (4,5) and (12,3), and then are asked to find the equation of the straight line that passes through the two points. These points as well as the straight line that connects them are represented in Figure 1.10.

Figure 1.10: A Straight Line Connecting Two Points



To obtain the equation of the straight line, it is necessary to first find the slope (m) as follows:

$m = \frac{\Delta y}{\Delta x} = \frac{5-3}{4-12} = \frac{2}{-8} = -\frac{1}{4}$

Once the slope is calculated, it is then possible to use one of two different approaches. One approach is to use the **point-slope form** of the equation, which is $y - y_0 = m(x - x_0)$. In this equation, (x_0, y_0) refers to any point on the line. Plugging in (4,5) as well as the slope leads to the following result:

$y - 5 = -\frac{1}{4}(x - 4)$

Solving for y yields the result below:
$y = -\frac{1}{4}x + 6$

Plugging in (12,3) and the slope would have led to the same result. The second approach involves the use of the **slope-intercept form** of the equation, which is y = mx+b. Plugging the slope and point (12,3) (or any other point on the line) into the equation yields the following result:
$3 = -\frac{1}{4}(12) + b$

Solving the equation for b yields the result below:
$b = 6$

We thus obtain the same result that we obtained when we used the point-slope form of the equation. That is:
$y = -\frac{1}{4}x + 6$

The mathematical concepts used in this textbook are relatively basic. The challenge is not so much in the mathematics but rather in the application of mathematical concepts to economic life. This review of basic mathematics is only intended to provide the reader with a helpful refresher as we begin our exploration of a wide variety of economic theories.

Following the Economic News

According to Harvard economist Raj Chetty, economics is a science.[11] In an opinion piece in *The New York Times,* Chetty acknowledged that the "root cause" of many disagreements in the field of economics is "our limited ability to run experiments." He pinpoints the practical and ethical costs of economic experiments as the problem. Nevertheless, he argues that "economists have recently begun to overcome these challenges by developing tools that approximate scientific experiments to obtain compelling answers to specific policy questions." He compares these advances in economics to "micro" advances in medicine (e.g., research on therapies for heart disease) even as "macro" questions (e.g., the determinants of health) remain the subject of considerable debate. As an example, he considers the claim that extending unemployment benefits increases unemployment rates by reducing workers' incentives to look for jobs. Several studies have compared unemployment rates in states that have extended unemployment benefits with those in states that have not extended benefits. He compares these studies to medical studies in which one group receives a treatment while another group does not receive it. That is, the control group does not receive the treatment. The studies uniformly found only a small positive impact on the amount of time people spent out of work when unemployment benefits are extended. Despite the analogy to medical experiments, the reader should keep in mind how much easier it is to control other factors when medical experiments are performed with subjects who are selected based on certain common characteristics. To what extent economics should be considered a science like medicine or physics is an issue that the reader should reflect on periodically while reading this book.

Summary of Key Points

1. Economics is best understood as an intellectual discourse about key social processes in our

society and how those processes affect human well-being.

2. Economics is a relatively young discipline that developed during the transition from feudalism to capitalism in Western Europe.
3. Adam Smith, David Ricardo, and Karl Marx were central figures in the evolving discourse of political economy in the eighteenth and nineteenth centuries.
4. Neoclassical economics dominates the discourse of economics today, and it is the perspective that is almost always presented to students in college economics courses as simply "economics."
5. Aside from the neoclassical school, all other schools of economic thought may be regarded as practicing heterodox economics.
6. The neoclassical entry point includes given resources, given technology, and given preferences, and its logic is one of unidirectional causality.
7. The distinction between positive economics and normative economics is central to neoclassical economics, and it is controversial.
8. The ceteris paribus assumption is the central assumption in all neoclassical economic models.
9. Independent variables are causes and dependent variables are effects in neoclassical models.
10. The cost function asserts that a positive relationship exists between output and total cost, other factors held constant.
11. The equation of a straight line passing through two points on the line may be found using either the point-slope form or the slope-intercept form.

List of Key Terms

Discourse

Economics

Feudal system

Market capitalism

Invisible hand

Neoclassical economics

Mainstream economics

Orthodox economics

Heterodox economics

Scientific method

Entry point

Logic

Unidirectional causality

Mutual causality

Positive statement

Normative statement

Economic models

Ceteris paribus assumption

Positively related variables

Negatively related variables

Independent variables

Dependent variables

Cost function

Total fixed cost (TFC)

Total variable cost (TVC)

Marginal cost

Economic forecasts

The tangent line method

Point-slope form

Slope-intercept form

Problems for Review

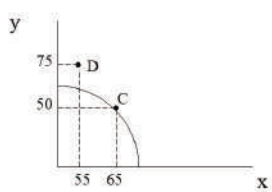
1. Suppose a reduction in income taxes occurs in the same year as a decrease in economic growth.
    - Is the relationship between these variables positive or negative?
    - Would you expect a positive relationship or a negative relationship between these variables?
    - If these variables are not related in the way that you were expecting, what might be a possible reason?
2. On a graph, plot the points (7,8) and (1,4).

3. Using the graph you drew in response to question 2, connect the points with a straight line.

    - Using point-slope form, write the equation of the straight line that passes through the points.
    - Using slope-intercept form, write the equation of the straight line that passes through the points.

4. Using the graph in Figure 1.11, use the tangent line method to determine the slope at point C. Assume the line also passes through point D when calculating the slope.

Figure 1.11: Problem 4



5. Using the data below, write the cost function using the form TC = TFC+(ΔTC/ΔQ)Q. Then draw the function on a graph with TC on the vertical axis and Q on the horizontal axis. Be

sure to label both axes correctly. Also identify the level of total fixed cost on the graph.

Quantity (Q)	**Total Cost (TC)**
90	70
180	100
270	130
360	160
450	190
540	220

Notes

1. For an advanced analysis of how these developments occurred, see Mirowski (1991).
2. The discussion of the neoclassical method described in this chapter draws upon ideas presented in *Contending Economic Theories* by Wolff and Resnick (2012).
3. Wolff and Resnick (2012), p. 37.
4. Ibid.
5. Ibid. Pp. 35-36.
6. Ibid. Pp. 35.
7. Ibid. Pp. 38.
8. Ibid. Pp. 36.
9. Robinson's quote is from Hunt (2002), p. 465.
10. Wolff and Resnick (2012), p. 7.
11. Chetty, Raj. "Yes, Economics Is a Science." *The New York Times*. October 20, 2013.

CHAPTER 2

WHAT IS A GOOD SOCIETY?

Goals and Objectives:

In this chapter, we will do the following:

1. *Describe* different perspectives of what constitutes a good society
2. *Define* the economic problem as understood within neoclassical economics
3. *Identify* the key prerequisites of economic efficiency in neoclassical economics
4. *Analyze* the production possibilities model of neoclassical economics
5. *Develop* two essential measures of opportunity cost that neoclassical economists use
6. *Explain* how economic efficiency is represented in the production possibilities model
7. *Apply* the production possibilities model to a number of different situations
8. *Build* a classification scheme for different economic systems

Competing Perspectives of What Constitutes a Good Society

The question of what constitutes a good society is one that has been discussed and debated for thousands of years. Many different answers have been offered, and no single answer is universally recognized as the best answer. The question was of interest in Ancient Greece. In *The Republic*, Plato described what he regarded to be the ideal society. Plato's ideal society was hardly a democracy. In fact, it contained a strict social hierarchy. For Plato, the best society was one in which the philosopher-kings ruled. This structure worked out well for Plato since he was a philosopher himself! The reason Plato regarded this feature to be an essential one in his ideal society is that philosophers were the wisest members of society. As a result, only they could be trusted to make decisions that would best serve the population. This class-based society also included a warrior class and a class of slaves, leading many to reject this notion of the ideal society.

Sir Thomas More offered another perspective on the question in the sixteenth century with the publication of his book *Utopia*. More approached the question from a Christian perspective in this famous work. In *Utopia,* More describes a fictional island named Utopia that is an ideal one from his perspective. The island of Utopia consists of several cities spread out across the island. The Utopians rotate living in the city and in the countryside. Their economy is based on communally owned property and the free distribution of necessities. They also have a short, six-hour workday and the least pleasant work is performed by slaves. The slaves are mostly criminals and prisoners of war. Freed from the toil of work, the Utopians were free most of the day to enjoy socializing with one another and learning about new subjects.

Mohandas Gandhi, the leader of the independence movement in India during the twentieth century, had a different perspective on what constitutes a good society. Gandhi advocated economic self-sufficiency, protectionism for local economies, and the avoidance of materialism.[1] The emphasis on local production and the resistance to long-distance trade in Gandhi's economic philosophy stem from the British legacy of colonialism in India. In the nineteenth century, a flood of British imports of cheap cloth helped bring about a rapid decline of the Indian textile industry. To this day, Indians have serious concerns about the possible consequences of unrestricted imports for domestic producers. An avoidance of materialism and excess consumption stems from a belief that spiritual values would suffer from an excessive devotion to the acquisition of material wealth. Gandhi also emphasized elevating the worst-off members of society in the pursuit of greater equality and a sense of brotherhood.[2]

Islamic economists hold a different vision of what constitutes a good society. Muslims believe that in the seventh century, God spoke to the Prophet Mohammad through the angel Gabriel.[3] These revelations were subsequently recorded in the Quran, and Mohammad's other declarations are documented in the Hadith.[4] These texts form the basis of the Islamic law code.[5] According to Islamic economists then, the ideal society is one that adheres strictly to Islamic law as set forth in the Quran and the Hadith. Whereas western nations abide by the principle of the strict separation of church and state, Islamic nations often have governments in which church and state are merged. These **theocratic governments** have the power to enforce compliance with Islamic law. Key features of Islamic economics include profit-sharing but a prohibition on the payment of interest, proper consumption that excludes the consumption of alcohol and pork, a wealth tax for redistribution to the poor, the avoidance of uncertainty (e.g., gambling), and a belief in universal brotherhood.[6]

Another thinker who advocated an ideal society in the nineteenth century was Karl Marx. Marx's advocacy of socialism (and ultimately communism) is somewhat unusual in that he did not specify with much precision how these superior societies were supposed to work. Instead, he concentrated his attention on the exploitative nature of market capitalism with the belief that his critique would ultimately lead to a revolutionary socialist transformation of human society. In Marx's theory, the capitalist class (the bourgeoisie) exploits the working class (the proletariat) within capitalist societies. According to Marx, the working class will eventually establish a society in which working people own the means of production in common and all class distinctions are abolished. Although Marx's description of the future socialist society is incomplete, he argued that workers would be compensated according to their work and would contribute to production as they are able. In the later communist society, workers would be compensated according to their need and would contribute to production as they are able.

The novelist and philosopher Ayn Rand is another thinker who offered a vision of the ideal society in the twentieth century. According to Rand, capitalism is the ideal society, as it is the only society, in her opinion, that protects human rights. For Rand, human rights are not granted by God, or society, or the state. Rather, rights derive from the very nature of human beings. These rights include the right to life and the right to property. That is, each individual human being has a right to do with his or her life or property whatever he or she wishes, provided he or she does not interfere with anyone else's rights. Therefore, no one has the right to take another person's life or property without first obtaining that person's consent. For that reason, the voluntary exchange of property is the only kind of transaction that respects the right to property. Market capitalism is thus the ideal society in Rand's view. Furthermore, the state has one function only and that is to protect the individual rights to life and property. If the state acts in a way that violates these rights, then it acts immorally.[7]

Alongside all these competing perspectives of the ideal society is the neoclassical perspective that the ideal society is the one that achieves efficiency. Simply put, the ideal society is the one that obtains the most advantages with the least use of resources. It may also be thought of as the avoidance of all waste. In the next section, we begin to explore the neoclassical concept of efficiency. As the reader will soon discover, neoclassical economists have a more complicated definition of efficiency than the brief definition offered in this section. Nevertheless, it should be clear that neoclassical economists possess their own concept of the good society.

The Economic Problem and the Scarcity of Resources

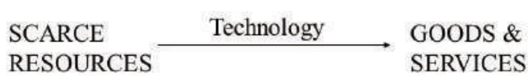
According to neoclassical economists, all human societies face the same basic problem, regardless of whether we are discussing a primitive tribal community or a modern, highly advanced industrialized economy.

The **economic problem**, as it is called, may be stated as follows:

- How can society best transform scarce and limited economic **resources** with its given state of **technology** into goods and services in the presence of unlimited and insatiable human **wants**?[8]

Figure 2.1 connects the key components of the economic problem.

Figure 2.1: The Economic Problem



At first glance, it might appear that the situation is entirely hopeless. The reason is that the resources used for production are limited and yet human wants are claimed to be limitless. It is, therefore, impossible to completely solve the economic problem. An infinite amount of production would be necessary to satisfy unlimited wants and that is simply impossible. It may be clear now why neoclassical economics is regarded as the dismal science. It begins with a problem facing all of humanity that can never be solved! What is the ideal society then if all societies confront this insoluble problem?

Before we investigate this issue further, it will be helpful to explore what neoclassical economists mean when they refer to resources. Resources, also referred to as **factors of production** or **inputs**, are classified into three different categories: land, labor, and capital. **Land** (T) refers to natural resources. That is, land includes all elements of production that are not previously passed through some production process, such as rivers, forests, and minerals that have not yet been extracted. **Capital** (K) is understood by neoclassical economists to mean something very different from what businesspeople often mean when they refer to capital. Businesspeople often have in mind a sum of money that is used to start a business. Occasionally, economists refer to **financial capital** when they wish to use the meaning that businesspeople assign to capital. Because economists are primarily interested in physical resources, however, they define capital as physical goods. Specifically, when neoclassical economists refer to capital, they have in mind goods that are used to produce other goods, including machinery, tools, and production plants. Similarly, the neoclassical use of the term **investment** (I) is rather different from the way the term is used in the financial services industry. Instead of referring to financial investments in stocks and bonds, neoclassical economists have in mind the growth of capital through time by means of production or purchase.

At this stage, it will be helpful to distinguish between two types of variables: stock variables and flow variables. A **stock variable** is a variable that is measured as of a point in time. A **flow variable**, on the other hand, is a variable that is measured per unit of time. For example, a household's **wealth** is a stock variable. That is, as of this moment, the household's total assets minus its debts can be measured in dollar terms (e.g., $250,000). The **income** of the household, on the other hand, is a flow variable. Household income can be measured in dollar terms, but the period must be specified. For example, it is not helpful to state that household income is $75,000. Is the income level $75,000 per year, per month, or per week? Similarly, capital is a stock variable, and investment is a flow variable. The capital stock may be 500 factories, but the amount of investment may be 15 factories per year.

Finally, **labor** (L) refers to the work that individuals perform. It may be measured as a stock variable or as a flow variable. For example, the size of the labor force might be measured in terms of the number of individual workers available for work at this moment, in which case it would be measured as a stock variable. On the other hand, it might be measured in terms of hours of employment each year, in which case it would be measured as a flow variable. Depending on the problem being considered, it may make sense to measure labor as a stock variable or as a flow variable. To avoid errors, it is important to know which type of variable has been chosen for the measurement of labor, capital, or land. Furthermore, regardless of how these variables are measured, it is crucial to remember that they are always regarded as scarce.

It should also be mentioned briefly that technology here means nothing more than the methods or techniques that are used to transform economic resources into goods and services. At a point in time, the knowledge of different production techniques is given or fixed. It is, therefore, an additional constraint that gives rise to the economic problem.

The Neoclassical Ideal: Economic Efficiency

It is now possible to consider what neoclassical economists consider to be the ideal society. The ideal society is one that achieves **economic efficiency**. The concept of economic efficiency in neoclassical theory has three key components. If any one of these three components is missing, then economic efficiency is not being achieved. First, the **full employment** of all resources must be achieved. That is, all land, labor, and capital must be fully used for production. Granted, it is necessary to consider efficiency in the use of resources over time. For example, if we use all our natural resources this year, then we will have nothing left in future years. As a result, it is necessary to distinguish between **static efficiency**, which refers to efficiency in the present moment, versus **dynamic efficiency**, which refers to efficiency over time. Although questions of sustainability are very important, our focus will be only on static efficiency in this chapter. That is, given the amount of resources our society has available in the present period, if all these resources are employed then full employment is achieved.

The second requirement for economic efficiency is **productive efficiency**, which refers to the least-cost or cost-minimizing method of production. It is possible that all resources are fully employed but not in the least cost manner. For example, suppose that for one day the heart surgeons trade places with the professional basketball players. Everyone continues to be fully employed but now they are employed in a way that is rather costly. The heart surgeries that take place that day will have a very high cost indeed, and the basketball games will probably not be especially well played.

The third requirement for economic efficiency is **allocative efficiency**, which refers to the most desirable mix of goods. That is, it is possible that all resources are fully employed and in the least-cost way but that the goods that are produced are not at all what people want to consume. For example, suppose that all that is produced in the economy is silly string. Everyone may have a job, but consumers will have no food, housing, or clothing to consume.

To sum up, for an economy to achieve economic efficiency, it must fully employ its resources in the least-cost manner and produce the combination of goods and services that consumers most desire. That is, full employment, productive efficiency, and allocative efficiency must all be achieved. The reader should note that an economy that achieves economic efficiency does not solve the economic problem. Instead, an economically efficient economy does the best that it possibly can in the face of the insoluble economic problem. The economically efficient economy is, therefore, the ideal society according to neoclassical economists. It is natural to ask which form of society will achieve this ideal outcome. We will return to this question towards the end of this chapter. In the next section, we will see how the different components of economic efficiency may be represented in an economic model.

The Production Possibilities Model

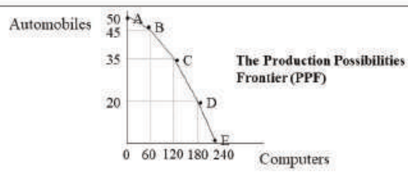
In this section, we will consider how to represent the economic problem within the context of an economic model called the **production possibilities model**. The model is a graphical depiction of the economic problem confronting any form of society. That is, because the economic problem is one that all societies face, this model is a completely general one in the sense that nothing is assumed about the type of economic system to which it applies. As with all economic models, it is necessary to identify the assumptions of the model. The model has several assumptions. It is assumed that the society only produces two goods, and they are measured in physical terms. The first assumption that the goods are measured in physical terms is significant because it is consistent with the notion that the model may apply to any form of society. Strictly speaking, the goods should not be measured in monetary terms because the economy being considered many not have money at all. It may be a barter economy or have some other mechanism of resource allocation and product distribution.

It is also assumed that the society possesses fixed stocks of land, labor, and capital, and a fixed production technology. The second and third assumptions that the resource stocks and production technology are given are central to the neoclassical entry point. Finally, it is assumed that the society is currently achieving full employment and productive efficiency. The assumption of full employment and productive efficiency is one requirement short of economic efficiency. That is, allocative efficiency is not assumed, and so at this stage, no assumptions are made about what society regards as the most desirable combination of goods. After the model is more developed, societal preferences will be introduced.[9]

Figure 2.2 shows an example of a **production possibilities schedule** with a corresponding **production possibilities frontier** (PPF).

Figure 2.2: A Production Possibilities Schedule

	Production Combinations				
Commodity	A	B	C	D	E
Computers	0	60	120	180	240
Automobiles	50	45	35	20	0



The table may be understood in the following way. A society may choose between production combination A, B, C, D, or E at any given time. For each production combination that it might choose, the society is able to produce the corresponding combination of computers and automobiles. For example, if society chooses production combination A, then it can produce both 0 computers and 50 automobiles. If it chooses production combination B, then it can produce both 60 computers and 45 automobiles, and so on. When each of these production combinations is plotted on a two-dimensional graph, a line can be drawn that connects them. This line is referred to as the production possibilities frontier (PPF). By assumption, each point on the PPF represents a combination of automobiles and computers that fully employs society's resources using the least-cost method of production (i.e., the best available production technology).

The production possibilities model teaches us two important lessons. The first lesson is that all societies face **tradeoffs**. One can easily observe from looking at the table in Figure 2.2 that an increase in the production of one of the goods necessarily leads to a reduction in the production of another good. Similarly, the negative slope of the PPF also indicates that society faces tradeoffs. As society moves along the curve from one production combination to another, a rise in the production of one good is only achieved due to a reduction in the production of another good. Another way of stating that all societies face tradeoffs is to state that all societies incur **opportunity costs**. The opportunity cost of an option X is the next best alternative Y that must be sacrificed to obtain X. Individuals incur opportunity costs all the time. The reader has chosen to read this book rather than take a nap, which might be the next best alternative. The opportunity cost of reading the book is the nap that cannot be taken. Similarly, societies incur opportunity costs. If society desires more computers, then it can only obtain additional computers by sacrificing the production of a certain number of automobiles. Due to full employment and productive efficiency, it is only possible to increase computer production by shifting resources away from the production of automobiles and towards computers. The loss of resources in the automobile sector is the reason that automobile production declines.

The second lesson that we learn from the production possibilities model is that all societies face *increasing* opportunity costs. That is, the opportunity cost of producing an additional unit of a good rises as more of it is produced. To really understand this concept, it is necessary to have a way of measuring opportunity cost. Otherwise, it does not make much sense to claim that opportunity costs are increasing. We have two ways of measuring opportunity cost.[10] The first way of measuring opportunity cost involves the use of something called **total opportunity cost**. The best way to understand total opportunity cost is to consider an example. Using the table from Figure 2.2, the reader will observe that if society produces 120 computers, then the entire amount of automobile production lost is 15 automobiles. That is, automobile production falls from 50 automobiles to 35 automobiles. In this case, the total opportunity cost of producing 120 computers is 15 automobiles. The reader should notice that the opportunity cost of a given amount of production of one good is always stated in terms of the other good that is lost. To consider a similar example, the total opportunity cost of 45 automobiles is 180 computers (because computer production falls from 240 to 60).

A second way of measuring opportunity cost, referred to as **marginal opportunity cost**, requires that we only look at small or marginal changes in the production of one good and the subsequent impact on the production of another good. For example, using the information from the table in Figure 2.2, we observe that an increase in computer production from 60 to 120 computers leads to a reduction in automobile production from 45 to 35 automobiles. That is, the marginal opportunity cost of the additional 60 computers is 10 automobiles. To consider a similar example, the marginal opportunity cost of increasing automobile production from 20 to 35 is 60 computers (because computer production falls from 180 to 120 computers).

When confronted with discrete as opposed to continuous data, it is sometimes necessary to calculate marginal opportunity cost using an approximation. For example, if we are interested in calculating the marginal opportunity cost of the 60th computer, then we do not have the information available in the table to answer the question with perfect precision. That is, we would need to see the movement from 59 to 60 computers and observe what happens to automobile production over this limited range. To approximate the marginal opportunity cost in this case, we will use the next lowest number in the table and divide the corresponding change in automobiles by the change in computer production as follows:

$$\frac{\Delta A}{\Delta C} = \frac{45A - 50A}{60C - 0C} = -\frac{5A}{60C} = \frac{-\frac{1}{12}A}{C}$$

According to this calculation, the marginal opportunity cost of the 60th computer is approximately 1/12 of an automobile. The perceptive reader might notice that ΔA/ΔC is also the slope of the PPF. The negative sign in our calculation indicates that the slope of the PPF is negative, which again implies that additional computer production has an opportunity cost (i.e., it leads to a loss of automobile production).

Now consider another example in which we attempt to calculate the marginal opportunity cost of producing the 120th computer. Again, we lack the information about how automobile production is affected when society increases production from 119 computers to 120 computers. Therefore, we use the next lowest number available, which in this case happens to be 60 computers. As before, we approximate the marginal opportunity cost as follows:

$$\frac{\Delta A}{\Delta C} = \frac{35A - 45A}{120C - 60C} = -\frac{10A}{60C} = \frac{-\frac{1}{6}A}{C}$$

Again, we observe that additional computer production carries an opportunity cost. However, we notice in this case that the opportunity cost has increased from 1/12 of an automobile to 1/6 of an automobile. That is, the marginal opportunity cost is increasing, which the reader will recall is the second lesson that we learn from the production possibilities model. Because neoclassical economists argue that all human societies experience the phenomenon of increasing marginal opportunity cost, they elevate it to the status of an economic law, which they call the **law of increasing opportunity cost**. Because the marginal opportunity cost is measured using the slope of the PPF, we can easily observe the operation of this law by noting how the slope becomes steeper (and thus the opportunity cost greater) as society moves along the PPF.

Of course, the reader might wonder why all human societies experience this phenomenon. After all, the numbers were simply assumed. They were not gathered during an empirical analysis of an actual economy. Why then do neoclassical economists regard it as self-evident that all human societies will experience this phenomenon? The reason relates to the underlying resources used in the production of the two goods. In general, societies possess **heterogeneous resources**. That is, among the resources societies use, some of the resources are better suited to the production of one good and other resources are better suited to the production of another good. If all resources were equally suited to the production of all goods, then society would possess **homogenous resources**. Because heterogeneous resources are the general rule, however, neoclassical economists assert that all human societies experience the phenomenon of increasing marginal opportunity cost.

To understand why the presence of heterogeneous resources leads to increasing opportunity cost, consider what happens when society begins at production combination A with 0 computers and 50 automobiles. As our society shifts resources from the automobile sector to the computer sector, computer production rises and automobile production falls. Notice, however, that automobile production only falls by 5 units when society moves to production combination B. The reason is that the first resources to be shifted to computer production will be the resources that are best suited to the production of computers and those that are least suited to the production of automobiles. When society moves to production combination C, computer production increases by 60 units again, but this time automobile production falls by 10 units. The reason for the larger decline in automobile production is that the resources shifted are now better suited to automobile production (and less suited to computer production) than those shifted previously. As a result, automobile production must fall by a larger amount to obtain the same rise in computer production that occurred previously. This pattern continues until automobile production must fall by a full 20 units (in the move from production combination D to E) to obtain a 60 unit rise in computer production. At this point, the best automobile resources must be shifted to computer production to obtain the 60 unit rise in computer production. It is the heterogeneity of the resource base that generates the pattern of rising marginal opportunity cost.

It should be noted that we could also approximate the marginal opportunity cost of an additional automobile in a similar manner. For example, if we wish to calculate the marginal opportunity cost of the 45th automobile, then we lack the information in the table that would allow us to calculate the answer exactly (i.e., the movement from 44 to 45 automobiles with the corresponding reduction in computer production). As a result, we can approximate the marginal opportunity cost of the 45th automobile as follows:

$$\frac{\Delta C}{\Delta A} = \frac{60C - 120C}{45A - 35A} = -\frac{60C}{10A} = \frac{-6C}{A}$$

Again, the negative sign indicates that additional automobile production carries an opportunity cost (i.e., lost computer production). The reader should notice, however, that ΔC/ΔA is the reciprocal of the slope of the PPF rather than the slope itself. Depending on which good we are considering, the marginal opportunity cost may be approximated using the slope or the reciprocal of the slope of the PPF. The important point to remember is that the *other* good is always placed in the numerator when carrying out this approximation. The reason the other good is placed in the numerator is that opportunity cost is always measured in terms of the lost amount of the other good and never in terms of the same good.

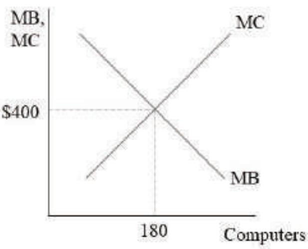
Sometimes the phenomenon of increasing marginal opportunity cost is described in terms of the **law of diminishing returns**. For example, consider the situation in Figure 2.2 where the society begins with 240 computers and 0 automobiles. As computer production declines in increments of 60 computers, the increases in automobile production become smaller and smaller. Society can boost automobile production a great deal initially, but the more it devotes land, labor, and capital to the production of automobiles, the more difficult it is to increase automobile production.

The Role of Societal Preferences in the Production Possibilities Model

Now that we have learned the two most important insights of the production possibilities model, we need to consider how economic efficiency is represented in the model. Because full employment and productive efficiency were assumed at the outset, any point on the PPF must satisfy these two conditions of economic efficiency. For society to achieve economic efficiency, however, allocative efficiency is another requirement that must be met. The reader will recall that for society to achieve allocative efficiency, the most desirable mix of goods must be produced. But what is the most desirable mix of goods for all members of society? We all have very different preferences and so sorting out what society's preferences are will not be an easy task.[11]

We can begin to think about how neoclassical economists address this question by dropping our earlier assumption that money values do not exist in this economy. If we consider an economy in which goods and resources have monetary values, then we can use something called **cost-benefit analysis** to determine the most desirable mix of goods for a given society.[12] Figure 2.3 shows how this analysis may be carried out using the computer industry from our previous example.

Figure 2.3: Marginal Benefit and Marginal Cost



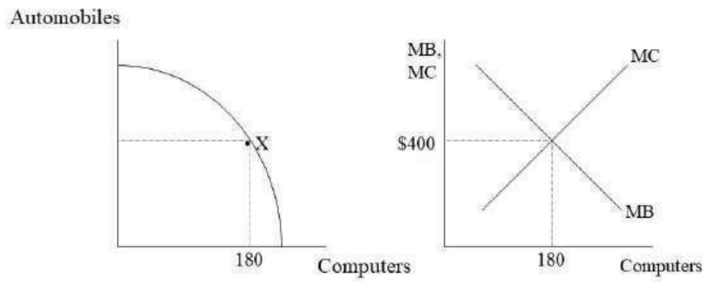
Unit	Marginal Benefit (MB)	Marginal Cost (MC)
0	$1100	$100
60	900	200
120	700	300
180	400	400
240	100	500

In Figure 2.3, it should be clear that as computer production rises, the marginal opportunity cost of production (here measured in dollar terms) increases. This increase in **marginal cost** (MC) is consistent with the law of increasing opportunity cost discussed previously. In addition, a new concept is introduced that we will refer to as **marginal benefit** (MB). The marginal benefit of a specific unit of a good represents the *maximum* dollar amount that members of society are willing to pay for that unit. In this case, when zero computers are produced, someone in society is willing to pay as much as $1100 for the first computer. By the time 60 computers are produced, the maximum amount someone is willing to pay is $900. The reason for the decline in marginal benefit is intuitive. As people consume more of a good, they experience less and less additional benefit from its consumption. Eventually, the marginal benefit of an additional unit declines to $100 once 240 computers are produced. What we observe in the table and the graph then are the **diminishing marginal benefit of consumption** and the **increasing marginal cost of production**.

When using cost-benefit analysis to determine the most desirable quantity of a good, it is only necessary to compare marginal benefit and marginal cost at each step. For example, when zero computers are produced, someone in society is willing to pay $1100 for the first computer, but the marginal cost to society is only $100. The production of the first computer clearly represents a net gain to society, and cost-benefit analysis indicates that it should be produced. When 60 computers are produced, a net gain to society exists from the production of another unit because the marginal benefit is $900 and the marginal cost is $200. The same holds true when 120 computers are produced. In that case, the marginal benefit of $700 exceeds the marginal cost of $300. It should be clear, however, that the **net marginal gain** (= marginal benefit minus marginal cost) is declining as more computers are produced. The marginal cost is rising due to the presence of heterogeneous resources, and the marginal benefit is declining as the population experiences a reduction in satisfaction from each additional unit consumed. Eventually, society reaches the point where marginal benefit and marginal cost are equal and net marginal gain is zero. When 180 computers are produced, the marginal benefit to society of an additional computer is $400, and this amount is exactly equal to the marginal cost of $400. In a sense, society just breaks even at this level of computer production. Although the net marginal gain of producing another computer is zero in this case, the computer can be produced without generating an inefficient outcome. Any additional production, however, will undermine economic efficiency. For example, if society produces 240 computers, then the marginal cost of $500 will exceed the marginal benefit of $100.

The general rule may be stated as follows. Society should produce a good up to the point where the marginal benefit of the last unit consumed equals the marginal cost of the last unit produced. If the marginal benefit exceeds the marginal cost of production, then society should produce more of the good. If the marginal benefit is below the marginal cost of production, then society should produce less of the good. Once the point is reached where marginal benefit equals marginal cost, then the allocatively efficient quantity of the good is being produced. This point occurs at the intersection of the MB and MC lines in the graph on the right in Figure 2.4.

Figure 2.4: The Condition for Allocative Efficiency



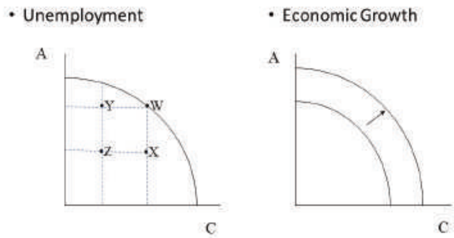
We can then identify in the graph on the left this quantity of computers. It is necessary to carry out a similar cost-benefit analysis for all goods produced in the economy. When all goods are produced such that marginal benefit equals marginal cost, then society has achieved economic efficiency. That is, society is not only producing on the PPF, indicating full employment and productive efficiency, it is also producing at the *optimal* (or best) point on the PPF from the standpoint of societal preferences, accounting for production costs. Again, we might return to the question of which form of society will bring about this result? Once again, the reader is asked to wait until the end of the chapter for the neoclassical answer to this question.

Relaxing the Assumptions of the Model

Up until now, we have continued to maintain the basic assumptions of the model. It is possible to relax these assumptions, however, and allow other variables to change. This violation of the ceteris paribus assumption is a useful way of understanding the role that the assumptions play. It also opens the door to applications of the model to historical and current events.

Let's begin by considering the assumption of full employment of resources. Suppose the economy is initially at point W in the graph on the left in Figure 2.5.

Figure 2.5: Relaxing the Assumptions of the Model



At point W, all resources are fully employed. Now suppose that some resources in the automobile sector become unemployed. Even though computer production will not be affected, automobile production will decline. This kind of change is reflected in the movement from point W to point X. Similarly, suppose that we begin at point W, but then resources in the computer industry become unemployed. Automobile production will not be affected, but computer production will decline. A change of this kind is reflected in the movement from point W to point Y. Finally, suppose that resources become unemployed in both sectors simultaneously. In that case, we would expect to see reductions in the production of both goods as reflected in the movement from point W to point Z.

The important point to notice here is that the economy moves inside its PPF when resources become unemployed. The PPF, however, does not move. The reason the PPF remains in the same place is that the resources to produce goods and services still exist. Hence, production combinations on the PPF are still possible. These production combinations are not produced, however, because some of the existing resources are idle.

It is also worth pointing out that the economy can move inside the PPF in a similar manner for a very different reason. Suppose that all resources are fully employed at point W, but then the automobile industry fails to use the least-cost method of production in that industry. In that case, the cost-minimizing production technology still exists, but it is not being used. As a result, automobile production declines below the maximum amount possible for that level of computer production, and the economy moves from point W to point X. Notice that resources are still fully employed at point X in this example. The reader should consider how similar failures to use the least-cost methods of production may lead to movements from point W to points Y or Z. Finally, such movements inside the PPF may involve a combination of the failure to fully employ society's resources and the failure to use the cost-minimizing production technology.

Another possible change worth considering involves a change in the amount of resources available to society. Suppose that the economy acquires additional stocks of land, labor, or capital. In that case, it will be possible to increase the production of both goods beyond combinations available on the PPF. The entire PPF will shift outwards in that case, and society will be able to produce more of both goods. Such an expansion of society's production capabilities is referred to as **economic growth**. This case is represented in the graph on the right in Figure 2.5. Similarly, a loss of resource stocks due to war, population decline, or sale will lead to an **economic contraction**, or an inward shift of the PPF.

The reason the PPF shifts inward in this case is that the resources themselves have been lost. This loss of resources is very different from the unemployment of existing resources. In the latter case, the resources still exist, but they are not employed. In the former case, the resources are no longer available at all and so society's production possibilities have changed.

Outward and inward shifts of the PPF may also result from a change in the existing state of technological knowledge. If society discovers new production technologies then it will be possible to produce production combinations that are beyond the initial PPF. That is, greater production of both goods becomes possible and economic growth occurs. On the other hand, a loss of production technology would lead to an economic contraction and an inward shift of the PPF. A loss of production technology may not seem like a very likely scenario, but it can occur if, for example, skilled workers begin to lose their skills due to long spells of unemployment. In this case, the knowledge of how to produce is lost, which is a very different scenario than the situation mentioned previously involving a failure to use the best available production technology.

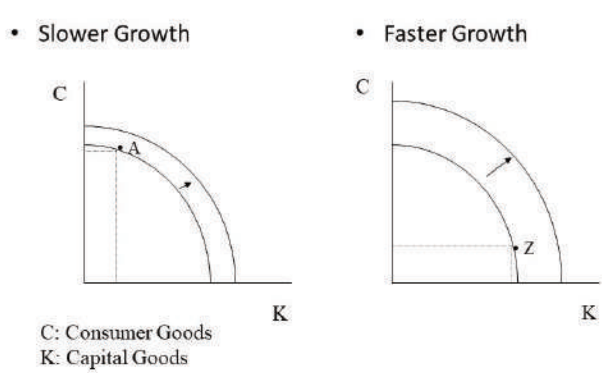
Applications of the Production Possibilities Model

We now have all the tools necessary to apply the production possibilities model to concrete situations. For example, consider the case of U.S. economic growth during the twentieth century. A number of factors led to the rapid expansion of the U.S. economy during that period, including population growth, growth of the nation's capital stock, and technological change. This situation is depicted in the graph on the right in Figure

2.5. Economic historians generally agree that, even though each of these factors contributed to the growth of the U.S. economy, technological change had the greatest impact and made possible the large increase in living standards that occurred during this period.

Another application involves the role of investment in the economy. The reader should recall that investment refers to the growth of the capital stock through time. To explore the implications of an increase in investment, we might begin by treating capital goods and consumer goods as our two goods in the production possibilities model. This approach departs a bit from our earlier presentation in that capital goods are simultaneously treated as a produced good and as a resource. Figure 2.6 shows an example of two different economies that have capital goods and consumer goods as their produced outputs.

Figure 2.6: The Role of Investment



In the graph on the left in Figure 2.6, one can see that the economy is producing a great deal of consumer goods and relatively few capital goods at point A. Because the economy invests little in new capital goods, the capital stock grows only slowly. Therefore, the PPF experiences a relatively small outward shift and the growth rate of the economy is relatively low. In the graph on the right, on the other hand, the economy produces a great deal of capital goods and relatively few consumer goods. Because it invests so much in new capital goods, the capital stock grows very rapidly and the production possibilities expand rapidly as well, as indicated by the outward shift of the PPF. It is sometimes argued that the Japanese economy grew much more rapidly than the U.S. economy during the post-World War II period precisely because the Japanese invested so much in new capital. The high saving rate among Japanese households made possible the accumulation of funds in banks, which then were loaned to Japanese corporations who used the funds for investment purposes. The so-called Japanese Miracle, which included extremely rapid annual growth rates, was the consequence. At the same time, U.S. consumers were eager to consume after having experienced considerable deprivation during the Great Depression and the subsequent period of wartime rationing. Because Americans spent a great deal of their income and saved little, the economic growth rate was lower than Japan's during this period. In general, the example illustrates that a tradeoff exists between present consumption and future consumption. Just as individuals might postpone present consumption to expand their future consumption possibilities, entire societies may also postpone present consumption and consume more in the future. Similarly, societies that consume a great deal today should not expect as much consumption in the future.

Another possibility is that an economy experiences **unbalanced growth**. This situation arises when one industry acquires additional resources or production technology without much of a change occurring in the other industry. When such a change occurs, the PPF shifts outwards, but the shift is skewed in favor of the industry that acquired significantly more resources or production technology. This situation is depicted in Figure 2.7.

Figure 2.7: Unbalanced Growth

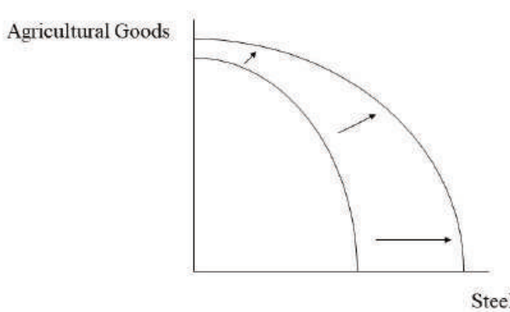


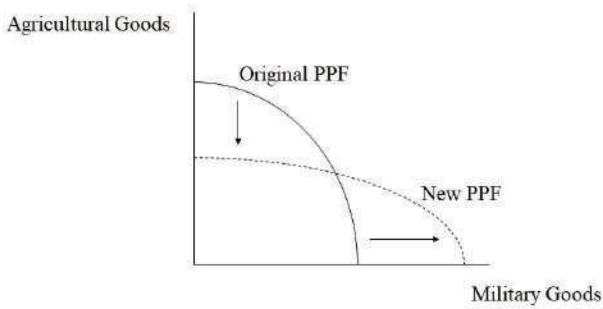
Figure 2.7 suggests that unbalanced growth has occurred in favor of the production of steel. This situation has arisen at times in the history of the steel industry when new steel furnaces were introduced. For example, the introduction of the open-hearth furnace that replaced Bessemer furnaces in the late nineteenth century and the

introduction of the basic oxygen furnace that replaced the open-hearth furnace in the mid-twentieth century led to growth in the steel industry (and in steel-consuming industries) that was more rapid than in other sectors. In general, it would be unusual for resource gains or technological advances to have a uniform impact across all industries leading to perfectly balanced economic growth. Therefore, when economic expansions occur, we should expect to observe unbalanced growth.

Of course, it is also possible for an economy to experience an **unbalanced contraction**. This situation arises when one industry loses resources or production technology without much of a loss occurring in another industry. When such a change occurs, the PPF shifts inwards, but the shift is skewed in a manner that negatively affects the industry that lost significantly more resources or production technology. This situation would also look like Figure 2.7 with a reversal of the arrows. In that case, the greater contraction would occur in the steel industry. These kinds of unbalanced contractions often occur in the case of natural disasters, such as tsunamis and hurricanes. For example, the 2004 tsunami, which created devastation in many nations in southeast Asia, mostly harmed the fishing and tourist industries while manufacturing and other industries located further from the coastline were left relatively unaffected.

It is also possible that one industry experiences economic growth while another industry experiences economic contraction. This situation is depicted in Figure 2.8.

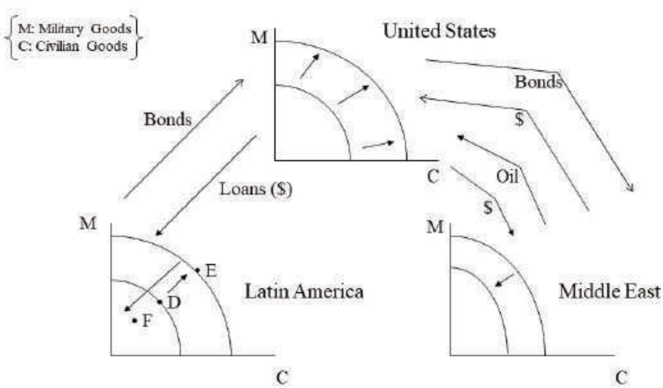
Figure 2.8: A Combination of Growth and Contraction



In Figure 2.8, the agricultural sector contracts at the same time that the military goods sector expands. This change might occur because the amount of arable land available diminishes due to soil erosion. At the same time, new production technologies in the production of military weaponry might lead to an expansion in that sector. In spite of this possibility, in many market economies, it is common to observe most industries contracting or expanding at the same time, albeit in an unbalanced way.

Let's consider a final example that demonstrates how the production possibilities model can link together several different regional economies. In the late 1970s and early 1980s, a phenomenon that came to be known as petrodollar recycling occurred. This process involved the United States purchasing large quantities of oil from Middle Eastern nations such as Saudi Arabia and Kuwait. The large accumulations of "petrodollars" that these nations received from the oil sales could not be profitably invested in their own economies given their low levels of economic development. As a result, the surplus dollars were recycled in the sense that they were used to purchase U.S. financial investments such as bonds. Figure 2.9 shows how U.S purchases of Middle East oil led to an expansion of production possibilities in the United States.

Figure 2.9: Historical Application: Petrodollar Recycling



At the same time, the loss of oil resources led to a contraction of production possibilities in the Middle East, particularly because a large amount of the petrodollars received were not used to obtain more resources or improved technology. Instead, as Figure 2.9 shows, the dollars were recycled through the purchase of U.S. financial assets. Additionally, U.S. banks loaned the petrodollars to Latin American nations, including Mexico, Brazil, and Argentina. These funds were badly needed due to high oil prices, but they were also used for an expansion of these nations' production possibilities, as shown in Figure 2.9. Once interest rates began to rise in the early 1980s in the United States, however, the burden of the rising debt became too great for these Latin American nations. Mexico defaulted on its debt in 1982. The ensuing panic led U.S. banks to restrict their lending to Latin American nations. The consequence was a sharp rise in unemployment, stagnant economic growth, and falling incomes. The economic boom in Latin America represented in the movement from D to E in Figure 2.9 was followed by a bust as represented in the movement from E to F. Because this process appeared to benefit the United States at the expense of Latin American nations, many critics have accused the U.S. of developing a neocolonial relationship with these nations.[13]

The Neoclassical Classification Scheme for Economic Systems

Even though the production possibilities model is a completely general model in that it may be applied to any economic system, neoclassical economists are primarily interested in understanding how market capitalist economies function. To better understand the differences between alternative economic systems, we will consider the neoclassical classification scheme for economic systems in this section.

According to neoclassical economists, economic systems have two primary elements.[14] The first element is the **form of property ownership**. Property ownership may take different forms, but the two forms that neoclassical economists consider to be the worthiest of our attention are **private ownership** and **state ownership**. Private ownership refers to ownership of the means of production and land by private citizens whereas state ownership refers to ownership of means of production and land by the government. The second element is the allocation mechanism that is used to allocate resources and to distribute goods and services. Different **allocation mechanisms** exist as well, but the two types that neoclassical economists consider to be the worthiest of our attention are **market allocation** and **centrally planned allocation**. Market allocation refers to the allocation of resources and the distribution of goods and services by means of markets. That is, resources and goods are transferred from buyers to sellers in the competitive marketplace where prices for these items are established by the free interaction of economic agents. Centrally planned allocation, on the other hand, involves the allocation of resources and the distribution of goods and services according to a central plan developed by government officials.

Because two forms of property ownership and two types of allocation mechanism have been identified, it is possible to name four possible economic systems based on how we mix and match these different elements. Figure 2.10 reveals that the neoclassical classification scheme gives rise to two types of market systems and two types of command systems.

Figure 2.10: A Neoclassical Classification Scheme for Economic Systems

		FORM OF OWNERSHIP	ALLOCATION MECHANISM
MARKET SYSTEMS	Market capitalism	*Private*	*Markets*
	Market Socialism	*State*	*Markets*
COMMAND SYSTEMS	Command Socialism	*State*	*Central Planning*
	Command Capitalism	*Private*	*Central Planning*

Market capitalism is the economic system that combines private ownership of the means of production and land with market allocation of resources and goods and services. Presently, most of the nations on Earth are either market capitalist economies, or they are in the process of transition towards market capitalist economies. The United States, Japan, Germany, the United Kingdom, Canada, Mexico, and so on, can all be regarded as market capitalist economies.

Not long ago, however, the defenders of market capitalism were engaged in a bitter dispute with the defenders of **command socialism**. In fact, these two economic systems were in direct competition with one another during the Cold War which ended with the dissolution of the Soviet Union in 1991. Command socialism merges state ownership of land and the means of production with a centrally planned allocation of resources. During the Cold War, many nations could be regarded as possessing command socialist economies, including the Soviet Union, the People's Republic of China, Cuba, and many other nations. Today, only Cuba and North Korea possess economies based on the traditional command socialist model.

Private ownership and market allocation frequently have been linked in an historical sense. Similarly, state ownership and the central allocation of resources have been linked throughout history. Figure 2.10 reveals, however, that two other economic systems are theoretically possible using this classification scheme. These **hybrid forms** or **cross forms**, as they are called, have existed historically, but they are much less common than market capitalism and command socialism.[15]

One of the hybrid forms is **market socialism**, which combines state ownership of the means of production and land with the market allocation of resources. Probably the best-known example of a market socialist economy is that of Yugoslavia from the early 1950s until the early 1980s. The Yugoslav economy was based on worker-managed enterprises that were "socially owned," but these enterprises operated in competitive markets.[16] The fact that Yugoslavia pursued a path that sharply deviated from the one Stalin wished to see coincided with considerable tension between the Yugoslav leadership and the Soviet leader.[17] In the 1980s, the economic system began to unravel due to ethnic and religious warfare that ultimately led to the fragmentation of the nation.[18]

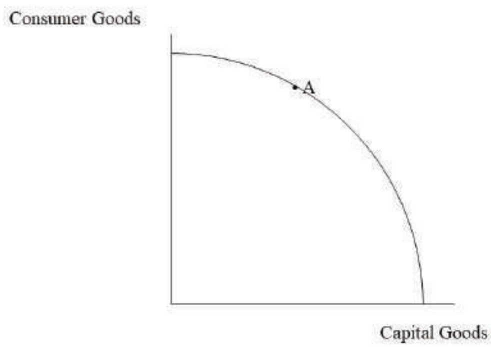
People often believe Sweden and other Scandinavian economies to be market socialist economies. This view is incorrect, however, if we rely on the neoclassical classification scheme. Sweden, for example, is characterized mainly by the market allocation of resources and goods as well as private ownership of land and the means of production. What leads many people to assume that Sweden is market socialist is that the government also engages in a significant amount of income redistribution. With high tax rates and extensive government programs to provide job training and income security, people regard it as socialist. Because the element of state ownership is not dominant, however, it cannot be correctly regarded as socialist using our schema. On the other hand, Sweden is sometimes regarded as a **social market economy**.[19] A social market economy is a market capitalist economy that is characterized by extensive income redistribution and a social safety net.[20] Sweden as well as other nations like it may, therefore, be correctly regarded as a special kind of market capitalist economy.

The second hybrid form is **command capitalism**. This economic system combines central planning and private ownership of the means of production and land. This economic system has existed at times, mainly during periods of wartime. The best-known example of a command capitalist economy is Nazi Germany.[21] Nazi Germany was characterized by the private ownership of the means of production, but government planning was used to direct the flow of resources to support the war effort. Major industrial producers, like Krupp, were privately owned and the profits from production were collected by the owners.[22] At the same time, they were not free to set production levels but instead accepted the orders of government officials.[23] Interestingly, the U.S. economy possessed significant elements of command capitalism during World War II.[24] Major American producers continued to be privately owned but wartime boards set production levels and prices to ensure victory in the war.

Of course, any given society will possess elements of both types of property ownership and both kinds of allocation mechanism. In other words, all economies are **mixed economies** to some degree, and we should not expect to observe any economic system in its pure form.[25] Which type of property ownership dominates and which type of allocation mechanism dominates in that society will determine how we choose to classify it using the classification scheme given here.

Now that we have a method of classifying different economic systems, we might ask whether any of these economic systems will achieve economic efficiency. In other words, we know that the ideal society according to neoclassical theory is the one that achieves economic efficiency, but which economic system is most likely to generate this outcome? If we can answer that question, then we will have a much clearer idea of what neoclassical economists regard as the good society, which is the main concern of this chapter. It turns out that neoclassical economists argue that market capitalism is the one economic system that can generate economic efficiency. That is, competition within markets between buyers and sellers of privately owned property will automatically lead, as if by an invisible hand, to the economically efficient outcome. The economy will end up both on the PPF and at the optimal point on the PPF accounting for societal preferences, as represented by point A in Figure 2.11.

Figure 2.11: Market Capitalism and the PPF



Unfortunately, the reader must wait for a complete explanation as to how market capitalism achieves this result. The neoclassical demonstration of this claim will be explored in detail in chapters 3 and 8.

Before leaving this subject, however, it might be worth mentioning that the widespread use of money in market capitalist societies is an important reason why neoclassical economists argue that such societies achieve economic efficiency. The use of money promotes efficiency because it solves a serious problem that exists in **barter economies** called the **double coincidence of wants problem**. Barter economies are economies in which goods are directly traded for one another without the mediating role of money to facilitate exchange. For example, suppose that Dave has 2 cows that he wishes to exchange for so many desks. In the absence of money, to successfully carry out this exchange, Dave must find someone who is willing to sell desks, but that same person must want to purchase cows. Hence, a double coincidence of wants problem exists in this case. That is, Dave must find a person willing to sell desks and that person must be willing to buy cows. If Dave finds such a person, then the price of the two cows in terms of desks must be negotiated as well. If money is present, on the other hand, then Dave only needs to sell his cows for money, and then he can use the money to buy desks from a seller of desks. Of course, the price of each good must still be negotiated, but it is no longer necessary that the seller of desks also be the buyer of cows. The double coincidence of wants problem is thus solved as soon as money enters the picture.

Money thus promotes efficiency because it reduces **transaction costs**, which are the costs associated with carrying out an exchange. Transaction costs are much higher in a barter economy than in a monetary economy because the time spent searching for a buyer who is simultaneously a seller is rather great in a barter economy. Furthermore, because money reduces transaction costs, it encourages people to rely more heavily on exchange to meet their needs. As a result, people are encouraged to specialize more than they otherwise would. Specialization increases efficiency, and so money promotes efficiency in this way as well.

A Marxian Classification Scheme for Social Formations

Before closing this chapter, we might consider an alternative to the neoclassical classification scheme for economic systems. Marxian economists use very different language to differentiate between economic systems. Even when the language overlaps, Marxists understand similar sounding economic systems in very different ways from their neoclassical counterparts.

To begin, Marxian economists generally refer to **social formations** rather than economic systems. The concept of a social formation is a very broad concept that has two major components. The first component is the **economic base** or **mode of production**. All the economic processes in society comprise the economic base. These processes relate specifically to how production is carried out and by whom. The second component is the **social superstructure**. The social superstructure consists of all the political, religious, cultural, and familial processes that reflect the underlying economic base.

Marxian economists recognize that the economic base and the social superstructure may each influence the other. For example, changes in the methods of production employed might influence political decision makers. Similarly, changes in religious beliefs and customs might influence how production is carried out. Although Marxian economists recognize multi-directional causality in their analyses of these concepts, traditionally they have given the greatest attention to the economic base. As Richard Wolff and Stephen Resnick argue, a twofold reason exists for this emphasis on the economic base and the conflict-ridden class relationships that often characterize it.[26] First, Marxists assert that class conflict is a greatly under-theorized aspect of human society. As a result, Marxists have frequently striven to place it at the center of their economic analyses. Second, Marxists believe that the neglect of class relationships in economic analysis has served as a barrier to the construction of an alternative society capable of eliminating class conflict. As a result, Marxian analyses often appear to give greater causal weight to the economic base even though this appearance stems primarily from a desire to call attention to the role of economic processes and class conflict in our society.

The economic base or mode of production also consists of two key parts. First, the **social forces of production** refer to the entire stock of accumulated knowledge of how to produce. All the production techniques, scientific expertise, and engineering capabilities of members of society fall into this category. The second major feature that defines a mode of production is the **social relations of production**. The social relations of production refer to all the patterns of interaction that exist between producers within a given society. One major difference that exists between modes of production is the presence or absence of **class exploitation**. Whenever a class other than the direct producers appropriates and distributes the products of the direct producers, then that mode of production is considered exploitative.[27]

Using the two key components of a mode of production, Marxian economists differentiate between several different modes of production.[28] The **primitive communist mode of production** consists of rather primitive social forces of production, including primarily hunting and gathering techniques. The social relations of production in the primitive communist mode of production involve production within relatively small familial or tribal groups that own the land and means of production in common. This kind of society does not involve class exploitation because not enough wealth is produced to create class distinctions and conflict.

The social forces of production in the **slave-based mode of production** consist mainly of agricultural production techniques. The social relations of production in the slave-based mode of production involve primarily the master/slave relationship. A small class of masters owns the land and means of production as well as a much larger class of slaves. The slaves carry on agricultural production, and the products that the slaves produce are appropriated and distributed by the master class. Because the production that occurs in slave societies is appropriated and distributed by someone other than the direct producers, slave-based forms of society are definitely exploitative.

The social forces of production in the **feudal mode of production** consist mainly of agricultural production methods as well. The social relations of production in the feudal mode of production involve primarily the lord/serf relationship. A small class of feudal lords within this mode of production employs a much larger class of serfs to work the land. The serfs are not exactly slaves, but they are not free either. Instead, serfs within the feudal mode of production are bound to the land by tradition, custom, and religious belief. The social hierarchy is accepted as part of the natural order. The serfs view it as their right to work the land by handing over a portion of their produce to the lord, and the lord views it as his right to appropriate and distribute this production. In return, the feudal lord provides the serfs with protection from hostile, outside forces. The feudal mode of production is also exploitative in nature because the direct producers do not directly appropriate and distribute the fruits of their labor.

The **capitalist mode of production** has been the mode of production of greatest interest to Marxian economists much as market capitalism has been the economic system of greatest interest to neoclassical economists. The social forces of production within the capitalist mode of production include large-scale machinery and mass production techniques employed within factories and large production plants. The social relations of production, on the other hand, involve production by a large class of wage laborers and the appropriation and distribution of their products by a small class of capitalists who employ them. The wage laborers are not slaves because the capitalists do not own them in the way that masters own slaves. Nevertheless, Marxian economists regard this mode of production as exploitative because the wage laborers do not appropriate or distribute the products of their labor. Instead, capitalists appropriate and distribute the products of wage workers because they own the means of production and the land.

The **socialist mode of production** is another mode of production of great interest to Marxian economists. According to Marxian economists, this mode of production will only come into being due to a revolutionary upheaval that abolishes capitalism and the distinction between capital and wage labor. Because this mode of production has not yet replaced capitalism, our description of it is based mostly on conjecture. Nevertheless, the social forces of production within the socialist mode of production are likely to include large-scale machinery and mass production techniques within complex enterprises. The social relations of production, on the other hand, will likely involve direct control of production enterprises by the direct producers. The direct producers will most likely own these enterprises in common, and compensation within enterprises will be linked directly to the duration and intensity of the work performed. The socialist mode of production is not exploitative because class distinctions will be abolished, and the direct producers will appropriate and distribute the fruits of their own labor.

The **communist mode of production** is another mode of production that has not yet come into existence. For Marxian economists, it represents the most advanced form of human society in which human beings have entered a harmonious relationship with nature. The communist mode production thus marks the end of history in the sense that the class struggle that underlies humanity's current economic problems permanently ends. Although the socialist mode of production abolishes the class struggle, some remnants of the capitalist mode of production will persist within the socialist mode of production. For example, only with the transition to the communist mode of production will compensation according to work performed be replaced with compensation according to need. This difference in compensation, which marks a difference in the social relations of production, is the primary difference between the socialist and communist modes of production. Otherwise, Marxian economists expect the two modes of production to be quite similar.

Figure 2.12 summarizes the differences between the modes of production that have been discussed in this section.

Figure 2.12: A Marxian Classification Scheme for Modes of Production

Mode of Production	Forces of Production	Social Relations of Production	Degree of Class Conflict
Primitive Communism	Hunting and gathering	No social classes	Non-exploitative
Slave-Based	Agriculture	Master/Slave	Exploitative
Feudal	Agriculture	Lord/Serf	Exploitative
Capitalist	Industrial mass production	Capitalist/Wage Laborer	Exploitative
Socialist	Industrial mass production	No social classes	Non-exploitative
Communist	Industrial mass production	No social classes	Non-exploitative

Marxian economists have identified other modes of production that are interesting, including the ancient mode of production and the Asiatic mode of production. The main purpose here is to introduce the reader to an alternative classification scheme for social formations (or modes of production to be more precise). These terms will be used repeatedly throughout the book and so it will be helpful to have a firm grasp on them moving forward.

Following the Economic News[29]

According to an article in *The Wall Street Journal*, many economists are concerned that cutbacks in public spending on capital projects in the European Union may negatively affect economic growth among member nations. For example, prior to the 2008 economic crisis, Ireland had planned to boost its investment in education infrastructure significantly. By 2011, however, concerns about the government's budget deficit led Irish policymakers to cut their plans for education by approximately half. One problem that Ireland is now experiencing is that it has a rapidly growing population but not enough school construction to accommodate the increase in the number of students. Even though more labor resources are being acquired, the failure to expand on society's capital resources will likely serve as a very serious barrier to the expansion of Ireland's production possibilities. In EU nations generally, public sector cutbacks have been occurring in transportation, housing, and education. Furthermore, efforts to cut government budget deficits have been advocated by those who wish to see salary cuts for public employees rather than cuts to infrastructure investment. In other words, these advocates of deficit reduction wish to see these economies produce relatively more capital goods and relatively fewer consumption goods. Salary cuts would discourage present consumption while maintaining investment in the future because infrastructure investment would remain high. Because it is politically much easier to cut spending on capital projects than it is to cut the salaries of public employees, it has been infrastructure investment that has been harmed the most. Defenders of the cuts argue that private sector investment will fill the void created by shrinking public sector investment. In any case, these changes pose serious challenges to economic growth in the nations of the European Union.

Summary of Key Points

1. Throughout history many different views of the good society have been put forward including, but not limited to, the Platonic, Christian,

Hindu, Muslim, Marxist, Randian, and neoclassical perspectives.

2. The neoclassical economic problem facing all human societies is how to transform scarce resources into goods and services to satisfy unlimited economic wants.
3. Land, labor, and capital are the three major resource categories in neoclassical economic theory.
4. Stock variables are measured as of a point in time and flow variables are measured per unit of time.
5. The neoclassical ideal of economic efficiency requires full employment of resources, productive efficiency, and allocative efficiency.
6. The production possibilities model offers a graphical representation of the economic problem and the conditions for economic efficiency.
7. The production possibilities model reveals that all societies incur opportunity costs and increasing opportunity costs due to the presence of heterogeneous resources.
8. Society achieves allocative efficiency when it equates the marginal benefit and the marginal cost of the last unit produced of each good or when net marginal gain is zero for each good.
9. When resources are unemployed or the cost-minimizing technology is not used, society moves inside its PPF. When resource stocks or the available technology changes, however, the PPF itself shifts.
10. The form of property ownership and the allocation mechanism determine the type of

economic system a society possesses in neoclassical economic theory.

11. In Marxian economic theory, a social formation depends on the mode of production and the social superstructure. The mode of production, in turn, depends on the social forces of production and the social relations of production.

List of Key Terms

Theocratic governments

Economic problem

Resources (factors of production or inputs)

Technology

Wants

Land

Capital

Financial capital

Investment

Stock variable

Flow variable

Wealth

Income

Labor

Economic efficiency

Full employment

Static efficiency

Dynamic efficiency

Productive efficiency

Allocative efficiency

Production possibilities model

Production possibilities schedule

Production possibilities frontier (PPF)

Tradeoffs

Opportunity costs

Total opportunity cost

Marginal opportunity cost

Law of increasing opportunity cost

Heterogeneous resources

Homogeneous resources

Law of diminishing returns

Cost-benefit analysis

Marginal cost (MC)

Marginal benefit (MB)

Diminishing marginal benefit of consumption

Increasing marginal cost of production

Net marginal gain

Economic growth

Economic contraction

Unbalanced growth

Unbalanced contraction

Form of property ownership

Private ownership

State ownership

Allocation mechanisms

Market allocation

Centrally planned allocation

Market capitalism

Command socialism

Hybrid forms

Cross forms

Market socialism

Social market economy

Command capitalism

Mixed economies

Barter economies

Double coincidence of wants problem

Transaction costs

Social formations

Economic base (or mode of production)

Social superstructure

Social forces of production

Social relations of production

Class exploitation

Primitive communist mode of production

Slave-based mode of production

Feudal mode of production

Capitalist mode of production

Socialist mode of production

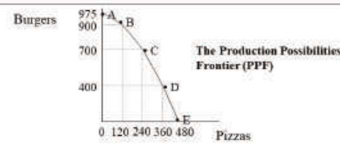
Communist mode of production

Problems for Review

1. Suppose society's production possibilities table and PPF are the ones shown in Figure 2.13.

Figure 2.13: Problem 1

Commodity	Production Combinations A	B	C	D	E
Pizzas	0	120	240	360	480
Burgers	975	900	700	400	0



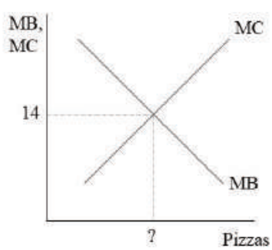
Answer the following questions:

- What is the total opportunity cost of producing 360 pizzas?
- What is the total opportunity cost of producing 900 burgers?
- What is the marginal opportunity cost of increasing pizza production from 120 to 240 pizzas?
- What is the marginal opportunity cost of increasing burger production from 400 to 700 burgers?
- What is the marginal opportunity cost of the 480th pizza?
- What is the marginal opportunity cost of the 700th burger?
- What is the marginal opportunity cost of producing the 900th burger?
- Considering your answers to the previous two questions, what happens to the marginal opportunity cost of producing burgers when burger production rises from 700 to 900

burgers? How would a neoclassical economist answer this question?

2. Using the graph and the table in Figure 2.14, can you fill in the missing pieces of information that are marked with a question mark?

Figure 2.14: Problem 2



Pizzas	Marginal Benefit (MB)	Marginal Cost (MC)
0	$22	$6
120	18	10
240	?	?
360	10	18
480	6	22

Explain the meaning of each number that you fill in as well as the significance of this row in the table.

Notes

1. For a nice summary of Gandhi's economic perspective, see Rosser and Rosser (2004), p. 90-91.
2. Ibid. p. 467.
3. Ibid. p. 94.

4. Ibid. 95.
5. Ibid. 95.
6. Ibid. 100-105. Rosser and Rosser provide a detailed overview of these Islamic economic principles.
7. See Rand (1966).
8. The reader should notice the emphasis on the neoclassical entry point in this statement of the economic problem. That is, given resources, given technology, and given wants or preferences are assigned a central place.
9. McConnell and Brue (1996), p. 25, postpone discussion of allocative efficiency until the details of the model are developed.
10. The use of the total opportunity cost and marginal opportunity cost measures may be found in McConnell and Brue's earlier editions. For example, see McConnell and Brue (1996). p. 26.
11. OpenStax College (2014), p. 36, asserts that the controversial nature of this question makes it a relevant question for sociologists, philosophers, and political scientists.
12. See Bade and Parkin (2013), p. 141-144, for one example of how this method is used to arrive at the allocatively efficient quantities of the two commodities.
13. For an overview of these issues and an explanation of how the debt crisis eliminated social welfare gains in the debtor countries, see Carrasco (2011).
14. Rosser and Rosser (2004), p. 6, identify six dimensions but agree that these two dimensions are the most important ones.
15. Ibid. p. 8. Rosser and Rosser also identify these four commonly recognized types of economic system.
16. Ibid. p. 389-390.
17. Ibid. p. 397.
18. Ibid. p. 389-390.

19. Ibid. p. 203.
20. Ibid. p. 203.
21. Ibid. p. 9.
22. Ibid. p. 9.
23. Ibid. p. 9.
24. Ibid. p. 9
25. Ibid. p. 9.
26. Wolff and Resnick (2012), p. 47.
27. Ibid. p. 27.
28. Ibid. p. 160-163. Wolff and Resnick describe these different modes of production as different forms of the fundamental class process.
29. Dalton, Matthew. "EU Infrastructure Cutbacks Worry Economists." *The Wall Street Journal*. July 20, 2014.

CHAPTER 3

THE NEOCLASSICAL THEORY OF SUPPLY AND DEMAND

Goals and Objectives:

In this chapter, we will do the following:

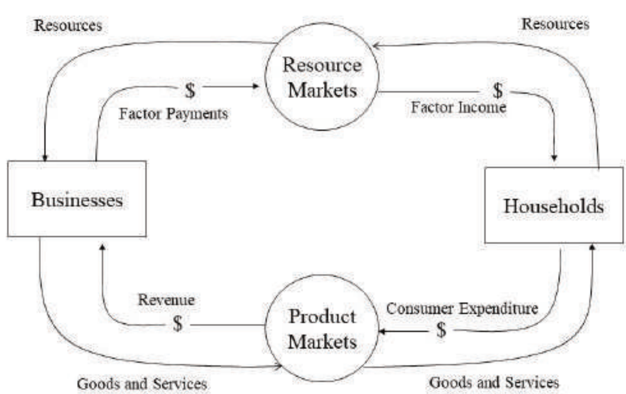
1. *Describe* the neoclassical circular flow model
2. *Define* supply, demand, and equilibrium
3. *Analyze* changes in equilibrium market outcomes
4. *Learn* how to apply the supply and demand model to actual events
5. *Define* and *calculate* the neoclassical measure of social welfare
6. *Investigate* situations in which markets fail to produce desirable results

The Neoclassical Circular Flow Model

Because neoclassical economists are primarily interested in understanding and defending market capitalist systems, they require a basic model that describes how the various parts of such economies are linked together. The **neoclassical circular flow model** achieves this goal. To understand this model, it is first necessary to define more carefully what we mean by a **market**. People often use the term to refer to a physical location where buying and selling occurs. For example, people will speak of "going to the market." Our definition of a market is much broader. It does refer to a coming together of buyers and sellers of a common good, service, or asset, but the buyers and sellers need not meet face to face. What is important is that they interact *in some way* through purchases and sales of a common item of exchange. For example, buyers and sellers may be spread out across the globe, interacting by computer only. Despite the distance between them, they participate in a market.

Given this definition of a market, we now turn to the neoclassical circular flow model represented in Figure 3.1.

Figure 3.1: The Neoclassical Circular Flow Model



As Figure 3.1 shows, the business sector and the household sector are linked together as a result of their interaction in two broad categories of markets. The first type of market is the **resource market**. Households own resources (or factors of production or inputs) that they sell to businesses. These resources include land, labor, and capital and so the major types of resource markets are the land, labor, and capital markets. The businesses use these resources to produce goods and services. In exchange for the resources, the businesses make factor payments to the households, which the households receive as factor income. For example, the payment for land generates rental income, the payment for labor generates wage income, and the payment for capital generates interest and profit income. The households use their factor income to purchase the goods and services that their resources produced in the second type of market referred to as the **product market**. The firms receive as revenue the consumer expenditure of the households.

Several important points need to be made about the circular flow model represented here. First, it should be noted that "real quantities" (i.e., resources and goods and services) flow in one direction in the diagram, and "nominal quantities" (i.e., dollar amounts) flow in the opposite direction.

Second, the model suggests that market economies are characterized by a strong element of social harmony. That is, buyers and sellers of resources and goods and services come together peacefully. Each gives up something he or she wants less than the thing obtained in exchange. Each market participant ends up better off at the end of each exchange than at the beginning. This world is the world of Adam Smith's Invisible Hand in which self-interested individuals are guided as if by an invisible hand to promote the social interest. The process will continue indefinitely as long as buyers and sellers are allowed to interact freely in the marketplace.

Third, it is easy to be misled by the diagram into believing that market economies possess a rigid class structure. That is, one might interpret the model to mean that business owners confront members of households in the markets. In fact, the entire population is represented in the household sector. Remember, the household sector includes all resource owners, including the owners of land, labor, and capital. Even a person who owns no capital or land still owns his or her own labor, even if it is unskilled labor.

Who then, is included in the business sector? The answer is that any members of the household sector may decide to start up businesses. They simply enter the resource markets and purchase resources from other resource owners. They then establish businesses and sell the finished goods and services to other members of the household sector. Business owners sell goods and services, but they also purchase goods and services for their own consumption as members of the household sector. Similarly, they purchase resources as business owners, but they may even sell their resources in the resource markets as members of the household sector. If they use their own resources to establish their businesses, then they need not sell to themselves although they will expect to receive enough revenue to compensate themselves for the use of these self-owned resources. The important point here is that the owners of businesses need not be owners of capital or land who then purchase labor. They may just as easily be owners of labor who purchase capital and land. In other words, no resource owner has a privileged position within the economic system. As Paul Samuelson famously remarked, in a perfectly competitive market, it does not matter whether capital hires labor or labor hires capital.[1] Therefore, no class structure exists, and no class conflict is inherent in the model.

The Buyers' Side of the Market

Now that we have a detailed overview of the interconnections that exist within a market economy, we can begin to analyze the neoclassical theory of how markets function. The neoclassical theory of supply and demand has three parts: **market demand**, **market supply**, and **market equilibrium**. Its purpose is to explain the two major outcomes of market competition: the *price* of the good, service or asset and the *quantity exchanged* of it. Before we combine the three parts of the model to show how they explain these two market outcomes, it will be helpful to provide a general overview of the method employed in the construction of the model. The method used is called **methodological individualism**. According to this approach, all social and economic outcomes are explained using individual actions. In terms of the supply and demand model, the starting point is individual buyers and individual sellers. Their actions are used to obtain the **individual demands** of the buyers and the **individual supplies** of the sellers. The individual demands are then aggregated (or summed up) to obtain market demand and the individual supplies are aggregated to obtain market supply. Finally, market demand and market supply are brought together to determine the market price and the market quantity exchanged. All these linkages are captured in Figure 3.2.

Figure 3.2: The Basic Structure of
the Neoclassical Supply and Demand Model

We will begin with the buyers' side of the market first and discuss individual demand. The individual demand for a good is the amount of a good that a buyer is willing and able to buy at each price during a given period. It is important to note that both willingness *and* ability to pay are central to the definition of demand. If an individual lacks the dollars to back up his or her willingness to pay, then his or her preferences do not count in the marketplace. Furthermore, according to neoclassical economists, all consumers are subject to an economic law known as the **law of demand**. The law of demand states the following:

Other things equal, a reduction in price causes the quantity demanded of a good to rise, and a rise in price causes the quantity demanded of a good to fall.

In other words, price and quantity demanded are inversely related, *ceteris paribus*. This negative relationship can be represented symbolically as follows:[2]

$$-\Delta P \Rightarrow +\Delta Q_D,\ \textit{ceteris paribus}$$

$$+\Delta P \Rightarrow -\Delta Q_D,\ \textit{ceteris paribus}$$

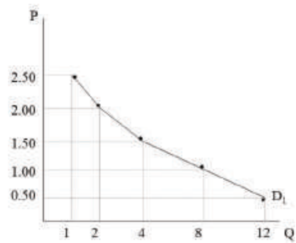
We represent the individual demand for a good using an individual demand schedule and an individual demand curve, as shown in Figure 3.3.

Figure 3.3: Representations of Demand

- An Individual Demand Schedule

Buyer 1's Demand for Apples	
Price per lb. ($/lb.)	Quantity demanded per month
2.50	1
2.00	2
1.50	4
1.00	8
0.50	12

- An Individual Demand Curve



The individual demand schedule shows clearly that the quantity demanded of apples (measured in pounds) per month rises as the price per pound of apples falls and vice versa. When the combinations of price and quantity demanded are plotted, we obtain a downward sloping individual demand curve. The downward slope implies that an inverse relationship exists between price and quantity demanded (i.e., that the law of demand holds).

At this stage, we have simply assumed that the law of demand holds, but it is important to understand *why* it is expected to hold. Neoclassical economists identify two reasons that we typically expect the individual demand curve to slope downwards:

1. As the price of apples declines, the consumer substitutes away from relatively higher priced goods, like bananas, and towards apples. This movement is known as the **substitution effect**. We might say that the consumer's *willingness* to purchase apples has increased because it is relatively cheaper.
2. As the price of apples declines, the consumer feels that his or her purchasing power has increased and so he or she decides to purchase more of all goods, including apples. This movement is known as the **income effect**. We might say that the consumer's *ability* to purchase apples has increased because his or her purchasing power has risen.

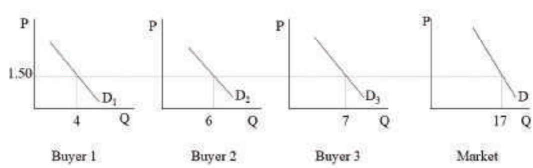
It is now possible to move from individual demand to market demand. Market demand refers to the quantity of a good that all buyers in a market are willing and able to purchase at each possible price during a given period. To obtain market demand, we must aggregate the individual demands of all individual buyers in the market. In terms of the individual demand schedules, we simply sum up the individual quantities demanded at each price to obtain the market quantity demanded at each price, as shown in Figure 3.4.

Figure 3.4: The Horizontal Summation of Individual Demands

Price Per lb.	Individual Quantities Demanded: First Buyer	Second Buyer	Third Buyer	Market Quantity Demanded Per Month
$2.50	1	2	3	6
$2.00	2	3	4	9
$1.50	4	6	7	17
$1.00	8	7	9	24
$0.50	12	10	11	33

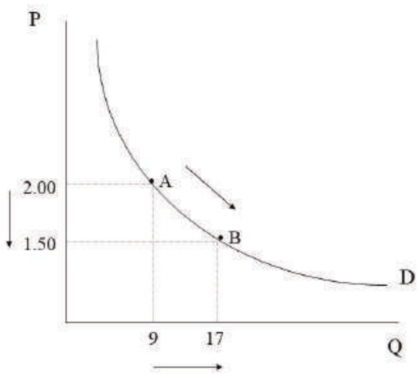
Graphically, this method of aggregation requires the summation of the horizontal distances from the vertical axis to the individual demand curve in each graph. For this reason, the method is referred to as **horizontal summation**. In Figure 3.5, this method is used to show the market quantity demanded at a price of $1.50. The other points on the market demand curve would be obtained in a similar fashion beginning with different prices.

Figure 3.5: The Market Demand Curve



Now that we have obtained the market demand curve, it is time to introduce a key terminological distinction in the discussion of demand. Neoclassical economists distinguish between a **change in quantity demanded** and a **change in demand**. Figure 3.6 shows the downward sloping market demand curve for apples. As the price falls from $2.00 per lb. to $1.50 per lb., the quantity demanded of apples rises from 9 lbs. to 17 lbs. per month.

Figure 3.6: A Change in Quantity Demanded



Such movements along the demand curve are referred to as changes in quantity demanded. In this case, the quantity demanded increases due to a price reduction. The quantity demanded also decreases when the price rises. It is important to note that the only factor that can cause a change in quantity demanded is a price change. In the case of the market demand curve, the price is the independent variable and the quantity demanded is the dependent variable. That is, the quantity demanded *depends* on the price. In mathematical terms, it is said that the quantity demanded is a function of the price of apples. That is:

$$Q_D = f(P)$$

Up until this point, we have acted as though the quantity demanded of a good depends only on the price charged for the good. Other factors are important to consumers as well, however, when they consider how much of a good to purchase. It is here that we begin to grasp the importance of the ceteris paribus assumption in the statement of the law of demand. The negative relationship between price and quantity demanded assumes that all other factors are held constant. If other factors change in addition to price, then this negative relationship may not hold. As a result, neoclassical economists assume that other factors are constant when tracing the demand curve. Nevertheless, neoclassical economists want to consider the possibility that other factors may change. These non-price determinants of demand include the following factors:

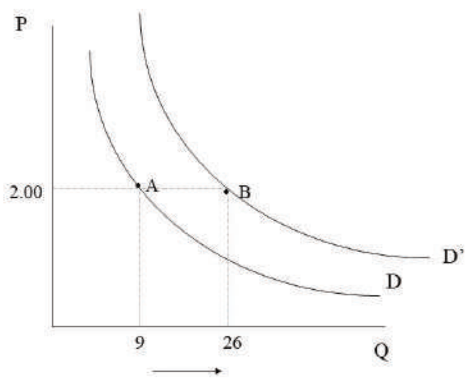
1. Consumer preferences or tastes
2. The number of buyers in the market
3. Consumers' incomes
4. The prices of related goods
5. Consumer expectations about future prices

This list of factors implies that the quantity demanded of a product is a function of the good's *own price* as well as all these other variables. That is:

$$Q_D = f(own\ price, \overline{preferences}, \overline{number\ of\ buyers}, \overline{incomes}, \overline{prices\ of\ related\ goods}, \overline{expectations})$$

The overbars indicate that these variables are held constant so that we can focus exclusively on the relationship between price and quantity demanded. When these variables change, however, the quantity demanded will be affected *at every price*. For example, Figure 3.7 shows what happens as Halloween approaches and the number of buyers in the market for apples increases.

Figure 3.7: Changes in Demand



The quantity demanded of apples at a price of $2/lb. rises from 9 lbs. per month to 26 lbs. per month. Similar increases in the quantity demanded occur at every possible price. As a result, the entire demand curve shifts to the right. Neoclassical economists refer to such a shift as an increase in demand (as opposed to an increase in quantity demanded). After Halloween, we might expect the quantity demanded of apples to fall at every price and the demand curve might shift to the left, back to its original position. Such a shift is referred to as a decrease in demand.

It is now possible to identify how changes in each of the non-price determinants of demand influence the demand curve. The case of a change in consumer preferences has already been discussed above in the context of a rise in the demand for apples:

Consumer preferences or tastes:

- A strengthening of preference ⇒ An increase in demand
- A weakening of preference ⇒ A decrease in demand

If more buyers enter a market, then the demand will also increase simply because it becomes necessary to aggregate a larger number of individual demands. If buyers exit the market, then fewer individual demands are aggregated and demand will decline.

The number of buyers in the market:

- An increase in the number of buyers ⇒ An increase in demand
- A decrease in the number of buyers ⇒ A decrease in demand

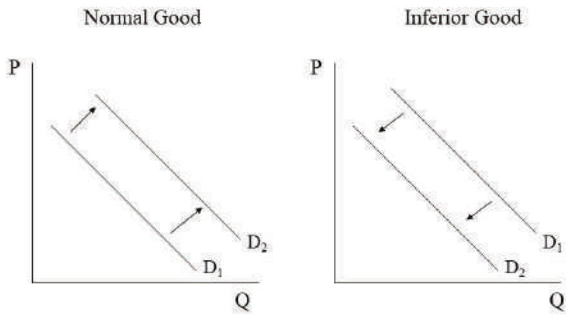
How changes in consumers' incomes affect demand depends on how consumers perceive the good in question. The demands of many goods are positively related to income. If consumers experience a rise in their incomes, they demand more of the good. If consumers experience a reduction in their incomes, they demand less of the good. These goods are called **normal goods**. The demands of some goods, however, are negatively related to income. When consumers experience a rise in their incomes, they demand less of the good. When consumers experience a drop in their incomes, they demand more of the good. These goods are called **inferior goods**. The reader might wonder why a consumer would purchase more of a good as his or her income falls. Examples of inferior goods include used goods and generic goods. As consumers experience a fall in their incomes, they substitute away from new goods and brand name products and instead purchase used goods and generic products. Hence, we have the following results:

Consumers' incomes:

- An increase in the incomes of consumers of a normal good ⇒ An increase in demand
- A decrease in the incomes of consumers of a normal good ⇒ A decrease in demand
- An increase in the incomes of consumers of an inferior good ⇒ A decrease in demand
- A decrease in the incomes of consumers of an inferior good ⇒ An increase in demand

Figure 3.8 shows how the market demand curves of normal and inferior goods are affected differently when incomes increase.

Figure 3.8: Changes in Demand due to an Increase in Consumers' Incomes



Changes in the prices of related goods may also cause changes in the demand for a good. If the price of a **substitute** for good A increases, for example, the demand for good A will rise, other things held constant. The reason is that consumers will substitute towards good A because it is now relatively cheaper. If the price of the substitute falls, then the demand for good A will fall as consumers substitute away from good A. For example, if the price of tea rises, we might expect the demand for coffee to rise as consumers switch from tea to coffee.

Alternatively, two goods may be **complements** in the sense that they are consumed together. If the price of a complement to good B increases, for example, then we would expect the demand for good B to fall, other things equal. If the price of a complement to good B falls, on the other hand, then we would expect the demand for good B to rise, other things equal. For example, if the price of peanut butter rises, then the demand for jelly will probably fall, other things equal. These results are summarized below:

Prices of related goods:

- An increase in the price of a substitute ⇒ An increase in demand
- A decrease in the price of a substitute ⇒ A decrease in demand
- An increase in the price of a complement ⇒ A decrease in demand
- A decrease in the price of a complement ⇒ An increase in demand

Finally, if consumers expect the price of a good to increase soon, then they frequently increase their demand for the good because they want to purchase the good before its price rises. For example, if consumers expect the price of gasoline to rise during Memorial Day weekend, they might demand more gasoline the weekend before that holiday to avoid paying the higher price. Alternatively, if the price is expected to fall in the future, then current demand will decline as consumers postpone their purchases in anticipation of a lower future price.[3]

Consumer expectations:

- A rise in the expected price of gasoline ⇒ An increase in demand
- A fall in the expected price of gasoline ⇒ A reduction in demand

The Sellers' Side of the Market

We now turn to the sellers' side of the market and discuss individual supply. (The reader will most likely notice a great many similarities between this discussion and the discussion of the buyers' side of the market!) The individual supply of a good is the amount of a good that a seller is willing and able to sell at each price during a given period. According to neoclassical economists, all sellers are subject to an economic law known as the **law of supply**. In particular, the law of supply states the following:

Other things equal, an increase in price causes the quantity supplied of a good to rise, and a decrease in price causes the quantity supplied of a good to fall.

In other words, price and quantity supplied are positively related, *ceteris paribus*. This positive relationship can be represented symbolically as follows:

$$+\Delta P \Rightarrow +\Delta Q_S,\ \textit{ceteris paribus}$$
$$-\Delta P \Rightarrow -\Delta Q_S,\ \textit{ceteris paribus}$$

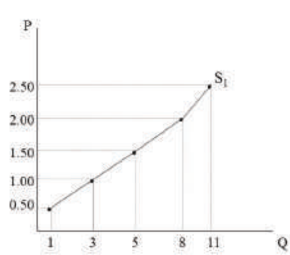
We represent the individual supply of a good using an individual supply schedule and an individual supply curve, as shown in Figure 3.9.

Figure 3.9: Representations of Supply

- An Individual Supply Schedule

Seller 1's Supply of Apples	
Price per lb. ($/lb.)	Quantity supplied per month
2.50	11
2.00	8
1.50	5
1.00	3
0.50	1

- An Individual Supply Curve



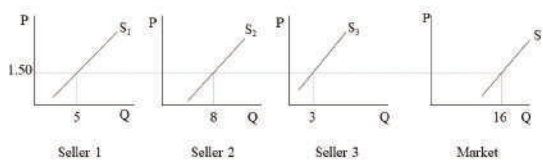
The individual supply schedule shows clearly that the quantity supplied of apples (measured in pounds) per month rises as the price per pound of apples rises and vice versa. When the combinations of price and quantity supplied are plotted, we obtain an upward sloping individual supply curve. The upward slope implies that a positive relationship exists between price and quantity supplied (i.e., that the law of supply holds).

At this stage, we have simply assumed that the law of supply holds, but it is important to understand *why* it is expected to hold. Neoclassical economists identify two reasons that we typically expect the individual supply curve to slope upwards:

1. As a firm increases production, the expansion puts upward pressure on per unit cost due to the scarcity of the firm's resources. The firm requires a higher price to cover the higher unit cost of production.
2. As the price rises, the firm finds it profitable to produce additional units of the product. It eventually stops due to the rising unit costs mentioned above.

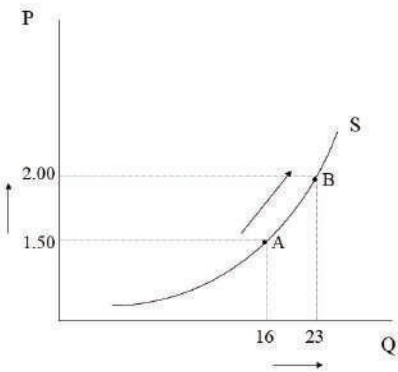
It is now possible to move from individual supply to market supply. Market supply refers to the quantity of a good that all sellers in a market are willing and able to sell at each possible price during a given period. To obtain market supply, we must aggregate the individual supplies of all individual sellers in the market. Just as in the case of demand, we use the method of horizontal summation. Graphically, this method of aggregation requires the summation of the horizontal distances from the vertical axis to the individual supply curve in each graph. In Figure 3.10, this method is used to show the market quantity supplied at a price of $1.50. The other points on the market supply curve would be obtained in a similar fashion beginning with different prices.

Figure 3.10: Horizontal Summation and the Market Supply Curve



Now that we have obtained the market supply curve, it is time to reintroduce the key terminological distinction that arose in the discussion of demand. Neoclassical economists distinguish between a **change in quantity supplied** and a **change in supply**. Figure 3.11 shows the upward sloping market supply curve for apples. As the price rises from $1.50 per lb. to $2.00 per lb., the quantity supplied of apples rises from 16 lbs. to 23 lbs. per month.

Figure 3.11: A Change in Quantity Supplied



Such movements along the supply curve are referred to as changes in quantity supplied. In this case, the quantity supplied increases due to a price increase. The quantity supplied also decreases when the price decreases. It is important to note that the only factor that can cause a change in quantity supplied is a price change. In the case of the market supply curve, the price is the independent variable and the quantity supplied is the dependent variable. That is, the quantity supplied *depends* on the price. In mathematical terms, it is said that the quantity supplied is a function of the price of apples. That is:

$$Q_S = f(P)$$

Up until this point, we have acted as though the quantity supplied of a good depends only on the price charged for the good. Other factors are important to sellers as well, however, when they consider how much of a good to sell. It is here that we should again recall the importance of the ceteris paribus assumption in the statement of the law of supply. The positive relationship between price and quantity supplied assumes that all other factors are held constant. If other factors change in addition to price, then this positive relationship may not hold. As a result, neoclassical economists assume that other factors are constant when tracing the supply curve. Nevertheless, neoclassical economists want to consider the possibility that other factors may change. These non-price determinants of supply include the following factors:

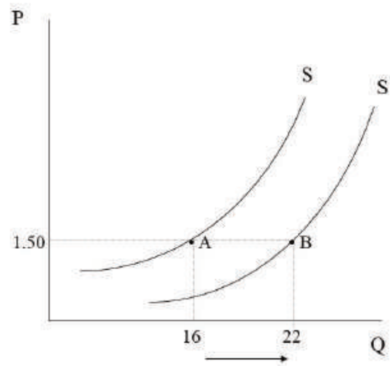
1. The number of sellers in the market
2. Factor prices
3. Production technology
4. Taxes and subsidies
5. The prices of related goods
6. The sellers' expectations about future prices

This list of factors implies that the quantity supplied of a product is a function of the good's *own price* as well as all these other variables. That is:

$$Q_S = f(own\ price,\ \overline{factor\ prices},\ \overline{number\ of\ sellers},\ \overline{technology},\ \overline{taxes},\ \overline{subsidies},\ \overline{prices\ of\ related\ goods},\ \overline{expectations})$$

The overbars indicate that these variables are held constant so that we can focus exclusively on the relationship between price and quantity supplied. When these variables change, however, the quantity supplied will be affected *at every price*. For example, Figure 3.12 shows what happens in late summer and fall when apples are in season and thus the number of sellers in the market for apples increases.

Figure 3.12: A Rise in Supply



The quantity supplied of apples at a price of $1.50/lb. rises from 16 lbs. per month to 22 lbs. per month. Similar increases in the quantity supplied occur at every possible price. As a result, the entire supply curve shifts to the right. Neoclassical economists refer to such a shift as an increase in supply (as opposed to an increase in quantity supplied). By late fall, we might expect the quantity supplied of apples to fall at every price and the supply curve might shift to the left, back to its original position. Such a shift is referred to as a decrease in supply.

It is now possible to identify how changes in each of the non-price determinants of supply influence the supply curve. The case of a change in the number of sellers has already been discussed above in the context of a rise in the supply of apples:

Number of sellers:

- An increase in the number of sellers ⇒ An increase in supply
- A decrease in the number of sellers ⇒ A decrease in supply

If factor prices change, then supply is affected. Factor prices (or resource prices or input prices) refer to the prices of land, labor, and capital. Specifically, if any of these factor prices rise causing rental payments, wage payments, or interest payments to rise, then firms will reduce their supply because their production costs have risen. Alternatively, if factor prices fall, then supply will rise because production costs are lower.

Factor prices:

- An increase in factor prices ⇒ A decrease in supply
- A decrease in factor prices ⇒ An increase in supply

Production technology is another important non-price determinant of supply. For example, when mass production methods were introduced in the automobile industry, production costs fell dramatically. Therefore, the supply of automobiles increased because firms could produce more at a given price. Although the implementation of an inferior technology is unlikely, it is theoretically possible and would lead to a reduction in supply as production costs rise.

Production technology:

- A technological advance ⇒ An increase in supply

- The implementation of an inferior technology ⇒ A decrease in supply

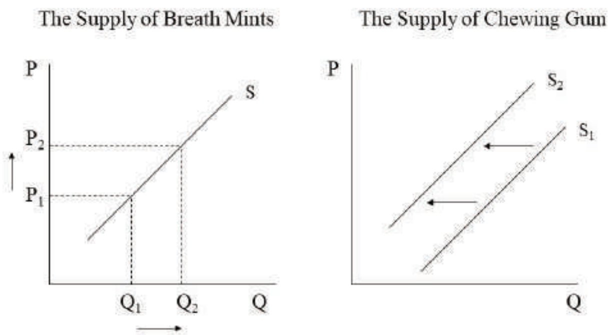
Changes in per unit **taxes** or **subsidies** on goods and services also influence product supply. For example, if the government increases the per unit tax on cigarettes, then production costs are higher and firms reduce the quantity supplied at every price. If the per unit tax on cigarettes falls, then firms increase their supply. Subsidies, on the other hand, have the opposite effect of taxes. Subsidies are cash payments that the government gives to sellers for each unit produced. If the government increases its subsidies for sellers, then production costs are effectively lower and product supply will rise. Alternatively, if the government cuts its subsidies to sellers, then product supply will fall.

Taxes and subsidies:

- An increase in taxes ⇒ A decrease in supply
- A decrease in taxes ⇒ An increase in supply
- An increase in subsidies ⇒ An increase in supply
- A decrease in subsidies ⇒ A decrease in supply

Prices of related goods sometimes also are important from the seller's perspective. For example, if the same firms produce two different goods and each firm possesses a fixed amount of resources, then a change in the price of one good will affect the supply of the other good. For example, suppose that the same firms produce breath mints and chewing gum. If the price of breath mints increases, then it becomes relatively more profitable to produce breath mints, other factors held constant. Because firms possess fixed resources, they will reallocate resources towards breath mints and away from chewing gum. As a result, the supply of chewing gum decreases as shown in Figure 3.13.

Figure 3.13: An Increase in the Price of a Related Good



Prices of related goods:

- An increase in the price of another good the firm produces ⇒ A decrease in supply
- A decrease in the price of another good the firm produces ⇒ An increase in supply

Finally, the expectations of producers about the future price may influence the supply of the product in a couple different ways.[4] For example, if the sellers expect the price to rise, then the sellers might expand supply now so that an increased quantity is present on the market when the price increase occurs. If they expect the price to fall, then supply might be reduced so that production is lower when the price decline occurs. On the other hand, a higher expected future price might lead sellers to store their current production and reduce supply. A lower expected future price might lead sellers to increase current supply so that they are able to sell before the price reduction.

Expectations about future prices:

- A rise in the expected price ⇒ An increase in supply (current expansion)
- A fall in the expected price ⇒ A reduction in supply (current contraction)
- A rise in the expected price ⇒ A decrease in supply (expansion of inventories)
- A fall in the expected price ⇒ An increase in supply (depletion of inventories)

Which outcome occurs depends on the circumstances related to the product. For example, durable goods can be stored for future sale whereas nondurable goods cannot be stored for long. Also, some price changes are expected soon whereas other price changes are expected in the distant future. The durability of the product and how soon the price change is expected are likely to interact to determine the impact on supply of an expected price change.

Equilibrium

We are almost to the point where we can combine the buyers' side and the sellers' side of the market to provide an explanation for the price and quantity exchanged of the product. Before we do so, we need to introduce a very important definition in neoclassical economics:

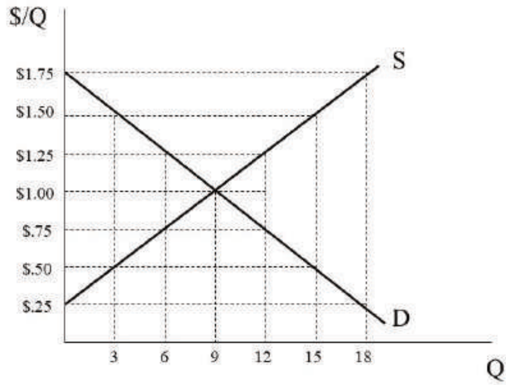
equilibrium. The equilibrium concept is central to the theorizing of all neoclassical economists. Its definition is simple:

An equilibrium state refers to a state from which there is no inherent tendency to change.

To understand this definition, consider a non-economic example.[5] For example, imagine a bowl that has a marble resting on the edge. If the marble slips into the bowl, it will move down along the inside edge eventually reaching the bottom and then rising again along the opposite inside edge. From that point, it will slip back down again where it will reach the bottom and then rise again on the other side. This back and forth motion will continue until finally, the marble finds its position at the bottom of the bowl. Is this state an equilibrium state? Yes, it is. Notice that I can shake the bowl and cause the marble to move. Nevertheless, it is an equilibrium state because the marble has no *inherent* tendency to move from this point. In other words, the definition of equilibrium does not preclude the possibility that external shocks might cause a disruption.

Neoclassical economists argue that market prices and quantities exchanged reach equilibrium levels much like the marble in the example. How is such a thing possible? The answer becomes clear when we bring together market supply and market demand as shown in Figure 3.14.

Figure 3.14: Competitive Market Equilibrium



To understand how the equilibrium price and quantity exchanged come about in this market, it is best to begin at a price that is *not* the equilibrium price. For example, suppose the price begins at $1.75 per unit. In that case, the quantity demanded is 0 units, but the quantity supplied is 18 units. Because sellers wish to sell 18 more units at this high price than buyers wish to buy, we say that an **excess supply,** or a **surplus,** of the good exists. Due to the surplus, sellers will feel pressure to cut price. Suppose that sellers cut the price to $1.50 per unit. As the price falls, the quantity demanded rises to 3 units due to the law of demand. Similarly, the quantity supplied falls to 15 units due to the law of supply. The result is a smaller surplus of 12 units (= 15 units – 3 units). The sellers will continue to feel pressure to cut price as long as a surplus exists. The reader should try to determine what the quantity demanded, the quantity supplied, and the surplus are at a price of $1.25 per unit.

Now consider what happens if the price begins at $0.25 per unit. In that case, the quantity supplied is 0 units, but the quantity demanded is 18 units. Because buyers wish to purchase 18 more units than are currently supplied, an **excess demand**, or a **shortage**, of the good exists. Due to the shortage, sellers will feel that they can increase the price. Suppose the sellers increase the price to $0.50 per unit. The quantity supplied will then rise to 3 units, according to the law of supply. Similarly, the quantity demanded will fall to 15 units due to the law of demand. The result is a smaller shortage equal to 12 units (= 15 units – 3 units). The sellers will continue to raise price if a shortage exists. The reader should try to determine the quantity demanded, the quantity supplied, and the amount of the shortage when the price rises to $0.75 per unit.

In the case of a shortage, the price will continue to rise until the price reaches $1.00 per unit. At this point, the quantity demanded is 9 units and the quantity supplied is 9 units. Therefore, neither a shortage nor a surplus exists in the market. As a result, sellers will not have an incentive to raise or lower the price. Every buyer who wishes to pay the current market price will be able to purchase the product, and every seller who wishes to sell at the current market price will be able to sell. All plans are consistent with one another. There is no inherent tendency for the price of $1.00 per unit or the quantity exchanged of 9 units to change. The outcome is, therefore, an equilibrium outcome. It is said that the market clears and so the equilibrium price and quantity are also referred to as the market-clearing price and quantity. These general results are summarized below:

- $Surplus : Q_S > Q_D \Rightarrow P \downarrow$

- $Shortage : Q_D > Q_S \Rightarrow P \uparrow$
- $Equilibrium : Q_D = Q_S \Rightarrow \overline{P}$

It is very important to note that the equilibrium condition states that *quantity supplied* equals *quantity demanded*. It is false to express the equilibrium condition in terms of supply and demand being equal (S = D). If supply equals demand, then the supply and demand curves are identical. That is, supply and demand refer to the curves themselves, whereas quantity supplied and quantity demanded both refer to specific points on the curves.

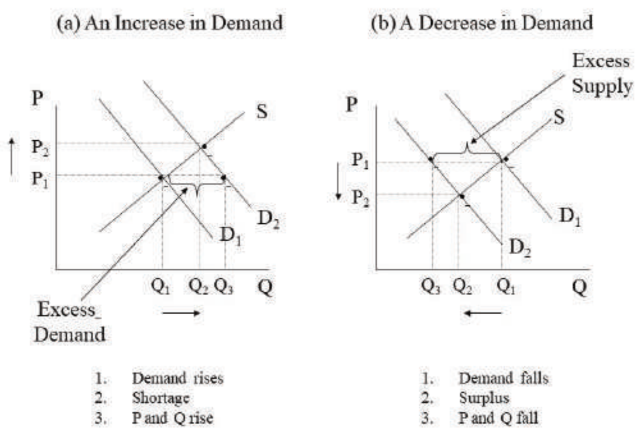
One other point should be made about the forces at work that bring about the equilibrium outcome. In the case of the marble moving towards its equilibrium level at the bottom of the bowl, the force at work is gravity. What is the analogous force at work in the market? The answer is *competition*. When economists refer to market forces, they have in mind competition between buyers and sellers.

Comparative Statics Analysis

It was mentioned earlier that equilibrium states can be disrupted by external shocks. The external shocks that can disrupt the equilibrium market price and quantity are changes in the non-price determinants of demand and supply that were discussed previously. When such changes occur, the supply curve or the demand curve will shift, bringing about a new equilibrium outcome. The analysis that allows us to compare these static equilibrium states is known as **comparative statics analysis**.

Let us consider changes in the non-price determinants of demand. For example, in the years following the Second World War, Americans were eager to purchase automobiles and household appliances. They had been deprived of many material things due to the Great Depression of the 1930s and the wartime rationing. With the postwar economic boom, many Americans experienced rising incomes, and they were ready to spend. Because these goods were normal goods, the demand for these goods rose. Figure 3.15 (a) shows the effect of an increase in demand on the market price and quantity exchanged for a good.

Figure 3.15: Changes in Demand

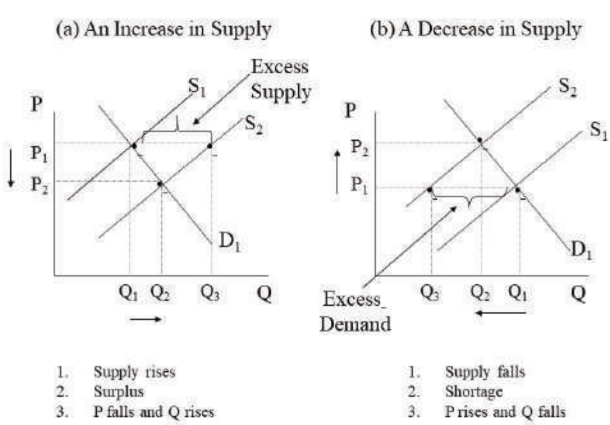


To understand how the movement from one equilibrium point to another occurs, it is useful to break up the movement into three steps. The first step is to identify whether supply or demand shifts and in which direction. The second step is to explain whether a surplus or a shortage exists *at the initial equilibrium price of P_1*. The third and final step is to explain the direction of the changes in price and quantity exchanged.

Figure 3.15 (b) shows what happens in the market for apples when the number of buyers decreases in late fall. The reduction in demand causes a surplus, which then causes the price and quantity exchanged to decline.

Similarly, Figure 3.16 (a) shows the effect of an increase in supply, which might occur due to a reduction in per unit taxes on cigarettes. Such a change creates a surplus and causes the price to fall and the quantity exchanged to rise. On the other hand, if a shortage of steelworkers causes wages to rise in the steel industry, then the rise in factor prices will lead to a reduction in the supply of steel as shown in Figure 3.16 (b).

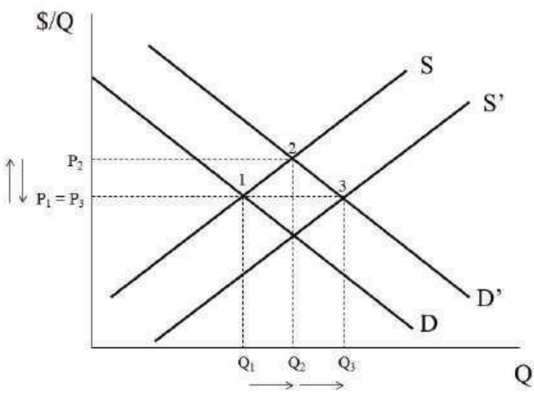
Figure 3.16: Changes in Supply



Simultaneous Shifts in Supply and Demand

Each of the cases discussed in the previous section involves a shift in supply or demand only. It is possible, however, that both supply and demand might shift simultaneously. For example, the supply of oil might increase due to the development of new production technologies while the demand for oil increases as developing nations become more industrialized. This case is illustrated in Figure 3.17.

Figure 3.17: A Simultaneous Rise in Supply and Demand



Based on the figure, it is clear that an increase in demand raises the equilibrium price and quantity exchanged. The equilibrium point moves from point 1 to point 2. The increase in supply, however, causes the equilibrium price to fall and the quantity exchanged to rise further. The final equilibrium point is at point 3. In this case, the reader can see that the quantity exchanged must rise because both factors contribute to a higher equilibrium quantity. On the other hand, the equilibrium price rises but then falls again. In this case, the price returns to its former level. One can easily imagine, however, a slightly larger increase in supply. Such a change would cause an overall reduction in price. Alternatively, a slightly smaller increase in supply would lead to an overall rise in price. The point is that unless we know the *magnitudes* of the shifts in supply and demand, we cannot determine whether price rises, falls, or stays the same overall. Our general conclusion, in this case, is that the quantity exchanged must rise, but the change in price is *indeterminate* (i.e., cannot be determined).

It turns out that in all four of the possible cases involving simultaneous shifts in supply and demand, the change in one of the variables (i.e., price or quantity) will be known for sure, and the change in the other variable will be indeterminate. The four possibilities are summarized below:

1. Supply and demand both rise: Q must rise but ΔP is indeterminate
2. Supply and demand both fall: Q must fall but ΔP is indeterminate
3. Supply rises and demand falls: P must fall but ΔQ is indeterminate
4. Supply falls and demand rises: P must rise but ΔQ is indeterminate

The reader should draw the graphs for cases 2-4, perhaps with the help of an instructor.

The Measurement of Social Welfare

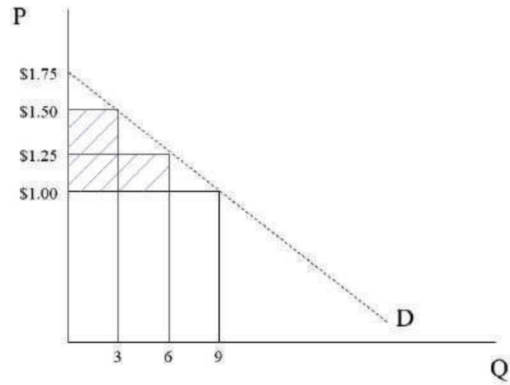
We have been discussing the determination of the equilibrium price and quantity exchanged using supply and demand analysis. Neoclassical economists are also interested in measuring how well-off consumers and producers are when the market is in equilibrium. To measure the well-being of consumers and producers, neoclassical economists use the concepts of **consumers' surplus** and **producers' surplus**, respectively. We will consider each concept in turn.

- Consumers' surplus refers to the difference between the maximum amount that consumers are willing and able to pay for a good and the amount they do pay for it.

For example, if you go to the mall to purchase a shirt, you may be willing and able to pay as much as $30 for the shirt. Because the price on the tag (the market price) is only $20, however, you enjoy the $10 difference in the sense that you acquire the shirt, *and* you retain the additional $10 that you would have been willing and able to pay. This surplus that the consumer enjoys is a measure of how well-off he or she is.

To calculate consumers' surplus for the entire market, we only need to recall the fact that the market demand curve represents the maximum prices that consumers are willing and able to pay for a good. Furthermore, because consumers pay the market price for the good, the consumers' surplus can always be represented as the area below the demand curve and above the current market price. Figure 3.18 shows how to approximate consumers' surplus by treating the demand curve as if it possesses a staircase shape. The shaded area represents this approximation of the consumers' surplus given the current market price of $1.00 per unit.

Figure 3.18: Consumers' Surplus



To calculate the consumers' surplus using this approach, it is only necessary to add up the shaded areas. We add the surplus of \$0.50 on each of the units from 0 to 3, the surplus of \$0.25 on each of the units from 3 to 6, and the surplus of \$0.00 on the units from 6 to 9.

$$CS = (1.50 - 1.00)(3 - 0) + (1.25 - 1.00)(6 - 3) + (1.00 - 1.00)(9 - 6)$$
$$= 1.50 + 0.75 + 0 = \$2.25$$

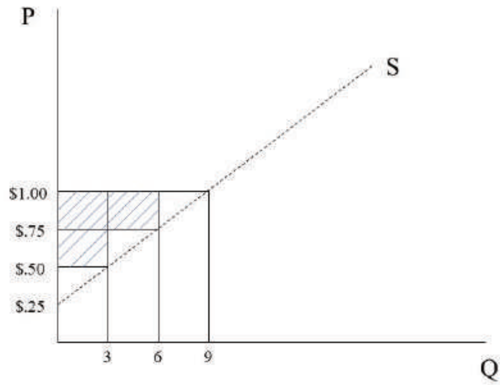
To calculate the consumers' surplus exactly, it would be necessary to include the entire area below the demand curve and above the price line. It should be clear that a larger area implies a greater welfare for consumers because the difference between the amount they are willing and able to pay and the amount they do pay is larger. The reader should also note that consumer expenditure is equal to \$9.00 (= \$1.00 per unit times 9 units).

We now turn to the related concept of producers' surplus. It is defined as follows:

- Producers' surplus refers to the difference between the amount that producers receive for a good and the minimum amount producers are willing and able to accept.

To calculate producers' surplus for the entire market, we only need to recall the fact that the market supply curve represents the minimum prices that producers are willing and able to accept for a good that they are selling. Furthermore, because producers receive the market price for the good, the producers' surplus can always be represented as the area above the supply curve and below the current market price. Figure 3.19 shows how to approximate producers' surplus by treating the supply curve as if it possesses a staircase shape. The shaded area represents this approximation of the producers' surplus given the current market price of $1.00 per unit.

Figure 3.19: Producers' Surplus



To calculate the producers' surplus using this approach, it is only necessary to add up the shaded areas. We add the surplus of $0.50 on each of the units from 0 to 3, the surplus of $0.25 on each of the units from 3 to 6, and the surplus of $0.00 on the units from 6 to 9.

$$PS = (1.00 - 0.50)(3 - 0) + (1.00 - 0.75)(6 - 3) + (1.00 - 1.00)(9 - 6)$$

$$= 1.50 + 0.75 + 0 = \$2.25$$

To calculate the producers' surplus exactly, it would be necessary to include the entire area below the price line and above the supply curve. It should be clear that a larger area implies a greater welfare for producers because the difference between the amount they receive and the minimum amount they are willing and able to receive is larger. The reader should also note that the revenue that producers receive is equal to $9.00 (= $1.00 per unit times 9 units). It should also be noted that the revenue that producers receive is equal to the expenditure of consumers. This result is perfectly consistent with the neoclassical circular flow diagram represented in Figure 3.1.

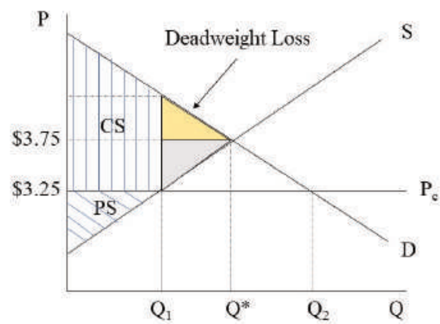
Although the concepts of consumers' surplus and producers' surplus have been represented here as being entirely consistent with the neoclassical worldview, we will consider two very important criticisms of these concepts in chapter 6.

The Economic Efficiency of Competitive Markets

We now come to an important conclusion that neoclassical economists draw. According to neoclassical economists, competitive markets achieve economic efficiency, as defined in chapter 2. In chapter 2, it is explained that a society achieves economic efficiency when it achieves productive efficiency, allocative efficiency, and full employment. Another way of thinking about economic efficiency is in terms of the maximization of consumers' surplus and producers' surplus. The **total surplus** (TS) is simply the sum of consumers' surplus and producers' surplus. When it is maximized, economic efficiency is achieved. Neoclassical economists argue that total surplus is maximized when the market is in equilibrium.

To understand their reasoning, we must take a slight detour into the analysis of government intervention in the marketplace. Suppose, for example, that the government imposes a **price ceiling** in the market for gasoline. A price ceiling is a maximum legal price that may be charged for a good or service. If the equilibrium price in the market for gasoline is $3.75 per gallon and the price ceiling is set at $3.25 per gallon, then the situation in Figure 3.20 is the result.

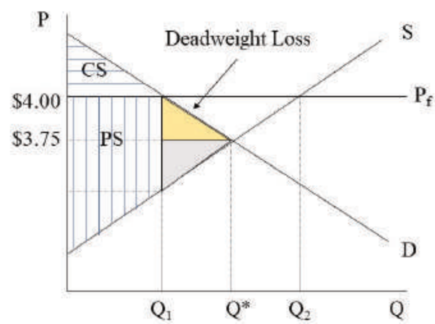
Figure 3.20: A Price Ceiling in the Gasoline Market



Because the price is prevented from rising above $3.25 per gallon, a shortage equal to $Q_2 - Q_1$ is the result. Furthermore, the shortage will persist indefinitely unless something else changes. The efficiency implications of the price ceiling are of even greater interest. Because only Q_1 units of output are supplied at the low price, the consumers' surplus and producers' surplus on units $Q^* - Q_1$ are simply lost. This loss of surplus is referred to as **deadweight loss**. Therefore, any price set below the equilibrium price will reduce the total surplus realized in the market.

On the other hand, the government might impose a **price floor** in the market for gasoline. A price floor is a minimum legal price that may be charged for a good or service. If the equilibrium price in the market for gasoline is \$3.75 per gallon and the price floor is set at \$4.00 per gallon, then the situation in Figure 3.21 is the result.

Figure 3.21: A Price Floor in the Gasoline Market



Because the price is prevented from falling below \$4.00 per gallon, an excess supply equal to $Q_2 - Q_1$ is the result. Furthermore, the excess supply will persist indefinitely unless something else changes. The efficiency implications of the price floor are of great interest just as in the case of the price ceiling. Because only Q_1 units of output are demanded at the high price, the consumers' surplus and producers' surplus on units $Q^* - Q_1$ are simply lost. As in the case of the price ceiling, deadweight loss is the result. Therefore, any price set above the equilibrium price will reduce the total surplus realized in the market.

For price floors to be **effective** or **binding**, they must be placed above the equilibrium price. A price floor set below the equilibrium price would not have any effect on the marketplace because price has no tendency to fall that low. Price will simply move to its equilibrium level. Similarly, price ceilings must be placed below the equilibrium price if they are to be binding. A price ceiling set above the equilibrium price would have no effect on the market participants. The price would simply fall to its equilibrium level because a price ceiling does not prohibit downward movements in price.

Maximum efficiency is, therefore, equivalent to maximization of total surplus and is achieved when the market reaches equilibrium. Any price charged other than the equilibrium price leads to deadweight loss. At the societal level, it is possible to understand how market equilibrium leads to economic efficiency in a different way. Specifically, economic efficiency requires full employment of resources, productive efficiency, and allocative efficiency. If each resource market clears, then the entire quantity supplied of each resource will also be demanded. As a result, full employment will be achieved. In addition, if we interpret the supply curve as representing the marginal cost (MC) of production and the demand curve as representing the marginal benefit (MB) of a good, then the equilibrium outcome implies that MB = MC, which is the condition for allocative efficiency. Unfortunately, we must wait until chapter 8 for a demonstration that market equilibrium leads to productive efficiency (or least-cost production).

The Possibility of Market Failure

It has been shown that free and unfettered markets lead to economic efficiency or the maximization of consumers' and producers' surplus. We now ask whether exceptions to this general rule can be found. When competition in the marketplace fails to bring about economic efficiency, we call this situation **market failure**. Several types of market failure have been identified. In this section, we briefly explore two broad types of market failure.

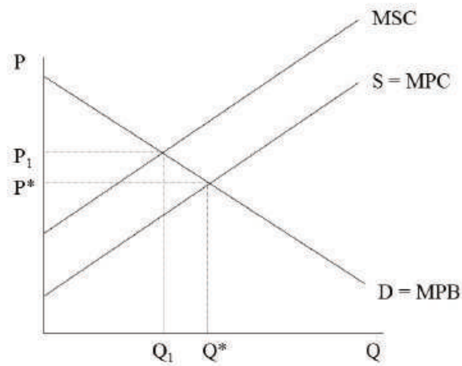
The first kind of market failure occurs with the production of **public goods**. Public goods are goods that have two key characteristics: non-rivalry and non-excludability. A good is non-rival if one individual's consumption of the good does not in any way reduce another individual's consumption of the good. The classic example is a lighthouse. Once the lighthouse is constructed and operational, all boats entering a harbor enjoy the benefit of the light, and one boat's consumption of the light does not diminish another boat's consumption of the light. By contrast, a private good, like a sandwich, is a rival good because one individual's consumption necessarily reduces another individual's consumption.

A lighthouse also has the characteristic of non-excludability. That is, it is not possible to exclude any one boat from consuming the service that the lighthouse provides once it is operating. As a result, if the lighthouse owner wishes to charge a fee for the use of the lighthouse, it will not be possible to do so. That is, a boat can consume the service without paying.

This situation creates a **free-rider effect** in the sense that each boat has an incentive to free ride on the efforts of other boat owners to pay for the construction of the lighthouse. Because all boat owners face this same situation, no one is willing to pay for the construction of the lighthouse, and it is not built even though the benefit to the boat owners outweighs the cost of construction. Therefore, the market fails to deliver the efficient amount of the good, and so it is a case of market failure. The case is frequently made that the government can enhance efficiency in these situations by imposing a tax on the boat owners to force the collection of the fees that would be needed to build the lighthouse.

The other kind of market failure that we will consider is referred to as an **externality**. Externalities are external effects of a market transaction on third parties that are not directly involved in the transaction. Externalities may be positive or negative. For example, a negative externality may occur when a steel company dumps waste into a nearby river. The waste imposes a cost on those living down the river. The price of steel, however, will not reflect this cost because it is not part of the private cost of production for the steel company. Neither the buyers nor the sellers of steel will experience the negative effects of this added cost of production. This situation is depicted in Figure 3.22 where MPC represents **marginal private cost**, MSC represents the **marginal social cost** of production, and MPB represents the **marginal private benefit**. The MSC includes both the MPC and the external cost of pollution. For this reason, the MSC is higher than the MPC.

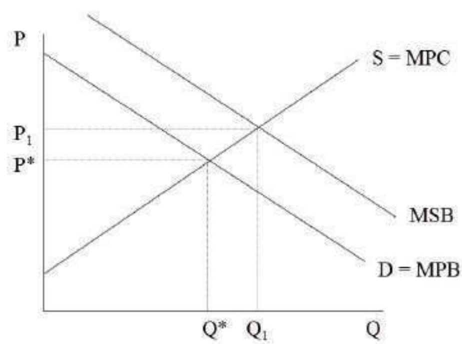
Figure 3.22: A Negative Externality in Production



In Figure 3.22, the market clears at P* and Q*. The efficient outcome, however, occurs where MPB intersects MSC at P_1 and Q_1. Because the efficient outcome occurs at a higher price and a lower quantity, the implication is that the good is *under-priced* and *over-produced* in a competitive market. The case might be made that a unit tax imposed on producers can be used to reduce supply and bring it into line with MSC. Such a tax is known as a **Pigouvian tax** after the economist A.C. Pigou.[6]

On the other hand, a positive externality may occur when consumption of a good leads to benefits for third parties that are not directly involved in the market transaction. For example, education is a good that is often argued to have positive spillover benefits in that other members of society benefit when people have more education rather than less. For example, a better educated population will have higher earning potential and so is less likely to resort to crime. Therefore, the **marginal social benefit** (MSB) exceeds the MPB in this case because the MSB includes the benefits that third parties receive in addition to the private benefits received by the direct consumers of education. This situation is depicted in Figure 3.23.

Figure 3.23: A Positive Externality in Consumption



In Figure 3.23, the market clears at P* and Q*. The efficient outcome, however, occurs where MSB intersects MPC at P_1 and Q_1. Because the efficient outcome occurs at a higher price and a higher quantity, the implication is that the good is *under-priced* and *under-produced* in a competitive market. The case might be made that a subsidy per unit given to consumers can be used to increase demand and bring it into line with MSB. Of course, it is difficult to know just how much of a tax or subsidy to impose in these cases of market failure. An error might lead to an even greater inefficiency than the one resulting in the case of competitive equilibrium.

Following the Economic News

The U.S. housing market has experienced a crisis in recent years with millions of homeowners unable to make their mortgage payments and losing their homes as a result. In many regions, homeowners have turned to renting due to the difficulties of obtaining mortgage loans. For example, this pattern has been observed in suburban Cook County in Illinois. According to U.S. Census Bureau data for the period 2007-2011, the demand for affordable housing in Cook County far outstripped the supply even though the number of affordable units increased. Researchers reported a shortage of nearly 48,000 affordable rental units due to the mismatch between the quantity supplied and the quantity demanded.[7] This case may be analyzed in terms of a simultaneous rise in supply and demand. As we have seen, the equilibrium quantity must rise. In this case, however, because demand has risen more than supply, we should expect rents (i.e., the price of rental units) to rise overall.

Summary of Key Points

1. The neoclassical circular flow model represents the linkages between the business sector and the household sector as harmonious and without class antagonisms.
2. The law of demand states, that other factors held constant, price and quantity demanded are inversely related.
3. A change in quantity demanded is due to changes in price only whereas a change in demand is due to changes in non-price determinants of demand.

4. The law of supply states, that other factors held constant, price and quantity supplied are positively related.
5. A change in quantity supplied is due to changes in price only whereas a change in supply is due to changes in non-price determinants of supply.
6. In market equilibrium, quantity supplied equals quantity demanded and neither a shortage nor a surplus exists.
7. Changes in the non-price determinants of demand and supply lead to changes in equilibrium prices and quantities exchanged.
8. Market equilibrium leads to the maximization of consumers' surplus and producers' surplus.
9. Binding price ceilings and binding price floors prevent the movement to equilibrium and reduce consumers' surplus and producers' surplus.
10. Market failures occur when markets fail to achieve economic efficiency.

List of Key Terms

Neoclassical circular flow model

Market

Resource market

Product market

Market demand

Market supply

Market equilibrium

Methodological individualism

Individual demands

Individual supplies

Law of demand

Substitution effect

Income effect

Horizontal summation

Change in quantity demanded

Change in demand

Normal goods

Inferior goods

Substitutes

Complements

Law of supply

Change in quantity supplied

Change in supply

Taxes

Subsidies

Equilibrium

Excess supply

Surplus

Excess demand

Shortage

Comparative statics analysis

Consumers' surplus

Producers' surplus

Total surplus

Price ceiling

Deadweight loss

Price floor

Effective or binding (price ceilings and price floors)

Market failure

Public goods

Free rider effect

Externalities

Marginal private cost (MPC)

Marginal social cost (MSC)

Marginal private benefit (MPB)

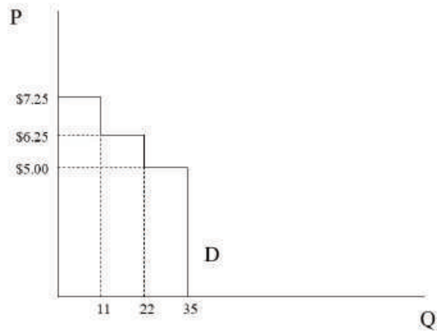
Pigouvian tax

Marginal social benefit (MSB)

Problems for Review

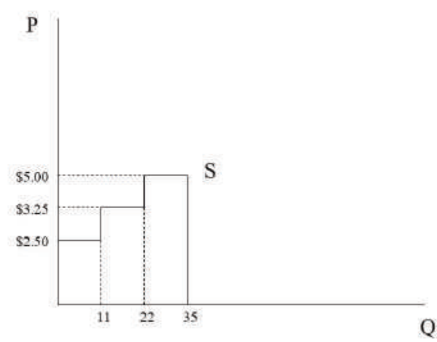
1. Suppose the price of peanut butter falls. If P_1 and Q_1 are the current equilibrium price and quantity of jelly, then how will the jelly market be affected by the drop in the price of peanut butter? Analyze this case in three steps.
2. Suppose the unit tax on gasoline is reduced. If P_1 and Q_1 are the current equilibrium price and quantity of gasoline, then how will the gasoline market be affected by the drop in the unit tax? Analyze this case in three steps.
3. Suppose consumers of an inferior good experience a rise in their incomes. If P_1 and Q_1 are the current equilibrium price and quantity of this good, then how will the market be affected by the rise in incomes? Analyze this case in three steps.
4. Suppose the demand rises for a good while the supply falls. Which conclusions can be drawn about the overall effect on equilibrium price and quantity exchanged? How would you draw the graph of this case?
5. Suppose the price of the good in Figure 3.24 is \$5 per unit. Calculate the consumers' surplus using the information in the graph.

Figure 3.24: Problem 5



6. Suppose the price of the good in the graph below is $5 per unit. Calculate the producers' surplus using the information in the graph.

Figure 3.25: Problem 6



Notes

1. Samuelson, Paul. 1957. “Wages and Interest: A Modern Dissection of Marxian Economic Models.” *American Economic Review*, 47: 884, 894.
2. This notation follows notation that Prof. David Ruccio used in his undergraduate economics principles class at the University of Notre Dame during my years as his teaching assistant in the early 2000s.
3. The reader should note that a movement along the demand curve does not occur in this case because the price of gasoline has not yet changed. It is only the expected future price of gasoline that has changed.
4. McConnell and Brue (2008), p. 52, recognize both possible outcomes whereas many authors of neoclassical textbooks only emphasize the case of current storage.
5. I believe that I first encountered this example in Prof. Kevin Quinn’s introductory economics class at BGSU.
6. For a nice summary of how externalities violate the neoclassical efficiency theorem and different approaches to resolving them, see Rosser and Rosser (2004), p. 32-36.
7. Podmolik, Mary Ellen. “Affordable Housing is Meager in Suburban Cook County.” *Chicago Tribune*. April 23, 2013.

Chapter 4

THE MARXIAN THEORY OF CLASS EXPLOITATION

Goals and Objectives:

In this chapter, we will do the following:

1. *Explore* the early history of Marxian economics
2. *Describe* a Marxian circular flow model
3. *Identify* the characteristics of commodities and money within the Marxian framework
4. *Explain* how to represent the circulation of commodities in Marxian economics
5. *Define* the concept of capital within Marxian economics
6. *Examine* the commodity labor-power and the determination of its value
7. *Illustrate* the division of the working day and the production of surplus value

The Early History of Marxian Economics

In this chapter, we will explore Marxian economics. As we learned in the last chapter, the neoclassical contribution to the discourse of economics places heavy emphasis on the roles of supply and demand in the marketplace as determinants of product and resource prices. Neoclassical economists also emphasize that free and competitive markets lead to economic efficiency. In contrast, Marxian economists place primary emphasis on human labor as the source of commodity value and assert that competitive capitalism leads to class exploitation. Before we explore the Marxian perspective in greater detail, we should briefly consider its origins.

In Chapter 1, we discussed the decline of feudalism and the transition to capitalism as it occurred in England and Europe during the eighteenth and nineteenth centuries. Indeed, the distinct discourse of political economy was born due to this transition. Classical political economists generally viewed the rise of capitalism as a positive development, but many critics of the new economic system also arose.

The most famous of these critics was Karl Marx (1818-1883). Marx was a man of many hats. He was a philosopher by training but was also an historian, an economist, a radical journalist, and a revolutionary political activist. At times, he even fancied himself a mathematician and a poet! He was born and raised in Germany, obtained his doctorate in philosophy, and participated in the revolutions of 1848 in Europe, during which time he wrote *The Communist Manifesto* with his collaborator and friend Frederick Engels. The *Manifesto* was essentially a call to workers throughout the world to resist the changes that were taking place in the global economy, which Marx considered to be detrimental to the well-being of workers. Despite his emphasis on the exploitation of workers, Marx did recognize that the economic transition was contributing to rapid economic development like nothing the world had ever seen. Marx was driven into exile due to his political activities. He moved to France and then Belgium, before settling with his family in England.

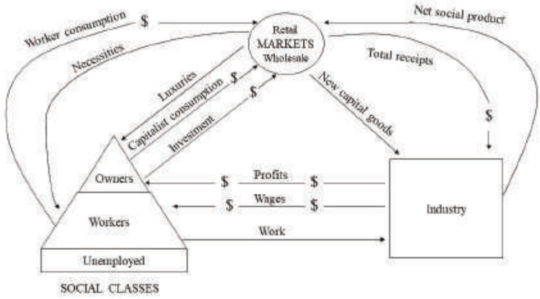
Once he settled in London, Marx began to concentrate on intellectual work, but political resistance was always a part of his life. In the late 1850s, Marx completed a work titled, *Grundrisse*, which means "Foundations." This work served as a rough draft for Marx's most famous intellectual work, *Das Kapital* or *Capital*. This three-volume work contains Marx's theory of how capitalism functions as well as his detailed critique of classical political economy. The first volume of *Capital*, published in 1867 and subtitled *The Process of Production of Capital*, is the only one that was published during Marx's lifetime. The second and third volumes, titled *The Process of Circulation of Capital* and *The Process of Capitalist Production as a Whole*, were published posthumously by Engels in 1885 and 1894, respectively. Another work titled *Theories of Surplus Value* was also published after Marx's death. It contains Marx's critical analysis of the history of economic thought and is sometimes regarded as the fourth volume of *Capital*. Because the foundation of Marx's theory of capitalism is developed in volume 1, it is our exclusive focus in this chapter.

Marx's theory of capitalism represents the merger of several complex bodies of thought.[1] Marx was inspired by the German philosopher G.W.F. Hegel. Hegel's theory of history was based on a kind of logic referred to as **dialectical idealism**. According to this manner of reasoning, all historical change is driven by the conflict between mutually antagonistic opposites.[2] Marx transformed this logic into one that emphasized material relations as the driving force in history rather than ideas as Hegel had done. It was later dubbed **dialectical materialism** for this reason. At the same time, Marx also borrowed the emphasis on the conflict between the **bourgeoisie** (capitalist class) and the **proletariat** (working class) from French political theorists.[3] Finally, he incorporated the theory that labor is the sole source of the value of commodities into his theory of capitalism.[4] Marx's theory of capitalism is somewhat unique in that it is an impressive blend of descriptive analysis and normative critique. Marx's concepts allow him to describe the workings of capitalism in detail, but the concepts inevitably lead to normative criticism. In this respect, his theory is sharply different from neoclassical theory, which aims to separate positive and normative analysis.

A Marxian Circular Flow Model

In the last chapter, we analyzed the neoclassical circular flow model, which shows how businesses and households interact through voluntary exchange in the product and factor markets. That model presents market relationships as harmonious, efficient, and consistent with Adam Smith's Invisible Hand of the market. We will now consider a **Marxian circular flow model** of a market capitalist economy that assigns a central place to the class struggle, which Marxian economists claim is the defining feature of market capitalism. A Marxian circular flow model, developed by Prof. Ed Nell, is shown in Figure 4.1.[5]

Figure 4.1: A Marxian Circular Flow Model



Source: Nell, Edward J. "The Revival of Political Economy." *Social Research* 39:1 (1972), 46. ©1972. The New School. Reprinted with permission of Johns Hopkins University Press.

The major difference between the Marxian circular flow model and the neoclassical circular flow model from Chapter 3 is that the Marxian model includes a well-defined social class hierarchy. That is, most of the members of this society are workers. A minority of the population consists of owners of the means of production, and this group has a higher place in the social hierarchy than workers. Beneath the workers is another large group of unemployed workers that does not contribute to the circular flow of the economy. In Marxian economics, this last group is referred to as the **reserve army of the unemployed**. The reader should recall that in the neoclassical model, all resource owners are lumped together in the household sector. Neoclassical economists do not find it necessary to distinguish between the owners of capital, labor, and land in their model.

Another major difference between the neoclassical and Marxian models is that in the Marxian model, the working class contributes to industry something that is fundamentally different from the capitalist class. That is, the working class contributes *work* to industry and the capitalist class contributes investment *funds* (or money capital), which are used to purchase new capital goods (i.e., means of production) to be used in industry. Each receives a different kind of income based on these different contributions to industry. Workers receive wages in return for their work, and capitalists receive profits as a return on their investments.

Another important difference between the two models is that in the Marxian model, capitalists and workers consume two different classes of goods and services. That is, the working class enters the retail markets to purchase **necessities** whereas the capitalist class enters the retail markets to purchase relatively more expensive **luxuries**. Because owners receive profit income, they can afford luxuries, like mansions, yachts, rare artwork, fine clothes, fancy automobiles, pricey jewelry, and catered meals. Because their wage incomes do not allow them to purchase luxuries, workers must be content with necessities, such as rental units, cheap clothing, used automobiles, inexpensive jewelry, and microwave dinners. It should also be noted that the sales of these goods and services generate monetary receipts for industry, which are used in the payment of wages and profits. In addition, only the **net social product** is sold in the wholesale and retail markets because a part of the product of industry is consumed during the production process. That is, industry uses up some of the means of production in the production of products and services.

Now that we have analyzed the neoclassical and Marxian circular flow models and considered the similarities and differences between the two models, the reader might like to know which of the two models is correct. That is, which model should the reader accept as the true representation of how market capitalist economies work? The answer is that we cannot settle this question for the reader. The reader must choose the model that he or she finds to be most convincing. The reader may even reject both models because neither model emphasizes those features of market economies that the reader finds to be most central. Remember that economics is a discourse with many competing opinions and viewpoints. If we identify one of these two models as containing the "Truth" about economics, then we have abandoned the notion that economics is a discourse and have returned to the conventional view that economics is a science that can provide final answers to descriptive questions.

To make the point in another way, we can see that neoclassical economists do not incorporate the potential for conflict between resource owners in their model. We can also see that Marxian economists do not incorporate the possibility of class mobility in their model. That is, can't a worker save enough money to join the class of capitalists? Each type of economist might be willing to admit these possibilities if asked, but the point is that they have excluded these possibilities because of all the infinite material that could be included in the models, they have *chosen* to include only those features that they consider to be the most important for understanding market capitalism. Consequently, their efforts to *describe* the functioning of economies have led them to *evaluate* that functioning from a normative standpoint at the same time. This inability to separate descriptive analysis from normative analysis is what makes economics a discourse at its core.

Commodities and Money

We can now begin to explore the theory of capital that Marx developed in volume 1 of *Capital*. Just like neoclassical economics, Marxian economics has its own language. Therefore, we will spend a great deal of time on its foundational concepts in this chapter just as we did when we explored the neoclassical theory of supply and demand. Once we have a solid grasp of the different components of Marxian theory, we can bring them together in a manner that provides a different conception of market capitalism than the one that neoclassical economists have developed.

Marx begins volume 1 of *Capital* by considering the form that **wealth** takes within the capitalist mode of production. He argues that wealth takes the form of an "immense collection of commodities." Because Marx regarded the **commodity** as the basic cell of capitalist wealth, it served as the starting point of his analysis. Marx next defines the commodity as consisting of two essential components. First, a commodity is a **use value**. That is, a commodity has a value in use. It serves some useful purpose for the user. We are surrounded by use values in our daily lives. For example, a chair allows one to sit rather than stand; a cell phone allows one to call or text one's friends; an automobile allows one to transfer one's self and one's belongings over long distances relatively quickly. This characteristic of a commodity is entirely *qualitative* in nature. To be a commodity, however, a thing must also be an **exchange value**. That is, the thing must also have a value in exchange. In other words, it must be possible to exchange the thing in the marketplace for so much of another commodity. This property of a commodity is entirely *quantitative* in nature. If a thing has both properties, it is a commodity.

If it has neither property, or only has one of these properties, then it is not a commodity.

In capitalist societies, commodities are everywhere. All the use values mentioned above are commodities because they are also exchange values. It should be noted that a thing need not be a material object to qualify as a commodity. For example, a haircut is something bought and sold in a marketplace. It is also useful to the person consuming it because that person walks away with a nice, neat haircut. Still, it is not a material object but because it is a use value and an exchange value, it is a commodity. All services sold to consumers may be regarded as commodities for this same reason. It should also be pointed out that some things do not qualify as commodities. In the United States, for example, it is illegal to own slaves. Although slaves have a use value, they do not have an exchange value due to the absence of an organized market. People are, therefore, not commodities in the United States, except where the practice continues as criminal activity. On the other hand, millions of people continue to live as slaves in parts of Africa and Asia. Where slaves are still bought and sold, people continue to exist as commodities.

Marx next turns to the question of what determines the exchange value of a commodity. Marx encounters two related questions when trying to explain the exchange value of a commodity:

1. How is it possible to equate two qualitatively different commodities through market exchange? Isn't this like saying apples and oranges are equal when apples are traded for

oranges in the marketplace even though they are clearly different things?

2. What determines the rate at which two commodities exchange? Why does 1 desk exchange for 5 books rather than 4 books or 6 books?

Marx solves this problem and answers these questions by arguing that a common element must exist within each commodity to make the equation of exchange possible. This common element, he argues, is human labor.[6] That is, even though the books and the desk are qualitatively different items, they both require human labor for their production. As a result, it is this hidden element that is equated in the process of exchange. Marx recognized, however, that the labor embodied in commodities has a dual character. On the one hand, the books and the desk each require a different type of labor for their production. The act of producing books is obviously very different from the act of producing a desk. These specific forms of labor Marx referred to as **concrete, private labor**. Because of the qualitative differences between these kinds of labor, we do not appear to have solved the problem of how qualitatively different commodities may be equated in exchange. On the other hand, Marx argues that the different production processes do have something in common. Each production process involves the generalized expenditure of human brains, muscles, nerves, hands, and so on. It is this homogenous human labor that is embodied in commodities and which is revealed when the exchange of two commodities takes place. Marx referred to this generalized human labor as **abstract, social labor**. It is the common element that makes it possible to equate the books and the desk in the process of exchange.

Furthermore, because labor time is a measurable concept, we also may use it to answer the second question above. That is, 5 books exchange for 1 desk because the 5 books and the 1 desk require the same amount of abstract, social labor for their production.

Marx's solution, therefore, suggests that a commodity will be more valuable when it requires more abstract, social labor for its production. A commodity will be considered less valuable when it requires less abstract, social labor for its production. Marx did recognize one potential problem with this explanation of exchange value. The solution appears to suggest that producers that are lazy and take longer to produce will produce commodities that are more valuable than efficient producers who complete production in a much shorter period. To resolve this problem with his theory, Marx concludes that only *socially necessary* labor contributes to the production of value. That is, the only labor that counts in exchange is the labor that is required under the conditions that are normal for a given society and that uses the average degree of skill and intensity prevalent in that society at a point in time. According to Marx's theory then, the **value** of a commodity depends directly on the amount of **socially necessary abstract labor time (SNALT)** required for its production. It follows that if only 10 hours of SNALT are required to produce a desk, then a producer that spends 15 hours of concrete, private labor on the production of a desk will have wasted 5 hours of labor time and only the 10 hours of SNALT will be realized in exchange. Similarly, a highly efficient producer that produces a desk with only 5 hours of concrete, private labor will be able to realize 10 hours of SNALT in exchange and so enjoys an extra bonus for this efficiency. Socially necessary abstract labor is, therefore, the *substance* of commodity value, and SNALT is its *measure*.[7] The claim that a commodity's value depends upon the SNALT required for its production, we will refer to as the **law of value**.

Another issue arises when we consider the concept of money. After all, when we go to the store to purchase a pair of shoes, we do not see on the price tag a specific amount of SNALT. The prices of commodities are stated in terms of money, not SNALT. As a result, Marx's next step is to explain why it is that commodity values are expressed in terms of money.

When Marx considers primitive economies in which commodity exchange only occurs sporadically and peripherally, he explains that value expressions can be represented with what he calls the **simple form of value** as shown in Figure 4.2.

Figure 4.2: The Simple Form of Value

20 yards of linen (20 hours of SNALT)	=	1 coat (20 hours of SNALT)

In this example, 20 yards of linen has its value expressed in the form of 1 coat because each requires 20 hours of SNALT for its production. This example is like the ones we have already considered.

As commodity exchange became a more widely adopted practice, generalized barter exchange economies developed. In these situations, value expressions may be represented with what Marx calls the **expanded form of value** as shown in Figure 4.3.

Figure 4.3: The Expanded Form of Value

20 yards of linen (20 hours of SNALT)	=	1 coat (20 hours of SNALT)	=	10 lb. of tea (20 hours of SNALT)	=	40 lb. of coffee (20 hours of SNALT)	=	etc.

In this case, each individual commodity has its value expressed in the form of *every* other individual commodity. Each commodity is taken in the appropriate quantity such that all quantities represent the same amount of SNALT. This form creates difficulties because value does not find its expression in a single form.

Due to the difficulties inherent in the expanded form of value, Marx explains that a single commodity is eventually set apart from all other commodities to serve as the **universal equivalent** for the expression of value.

The **general form of value** represents value expressions for this case as shown in Figure 4.4.

Figure 4.4: The General Form of Value

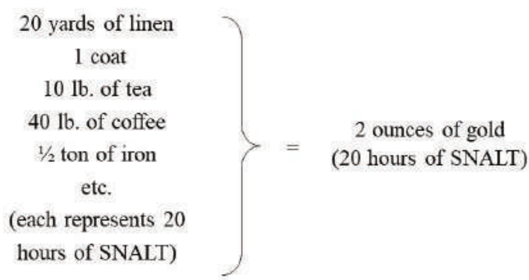
1 coat
10 lb. of tea
40 lb. of coffee
½ ton of iron
etc.
(each represents 20 hours of SNALT)

= 20 yards of linen (20 hours of SNALT)

In this case, linen is set apart from all other commodities and serves to express their individual values. All commodity values are now expressed in the form of linen. Again, the SNALT embodied in each commodity and the linen makes the equality between them and the linen possible.

Finally, Marx recognized that not just any commodity has served as the universal equivalent throughout history. Certain commodities have been more widely used due to their physical properties that make them likely choices for the universal equivalent. Precious metals, like gold and silver, are durable and easily divisible and so they are natural choices for the money commodity. In volume 1 of *Capital*, Marx assumes that gold obtains a social monopoly of the expression of value, which was generally the case in the nineteenth century. The **money form of value** thus represents value expressions for this case as shown in Figure 4.5.

Figure 4.5: The Money Form of Value



In this case, each commodity finds its value expression in the form of an amount of gold containing the same amount of SNALT for its production.

It is now possible to represent a commodity's value (or price) in terms of money as shown in Figure 4.6.

Figure 4.6: The Price-Form

20 yards of linen (20 hours of SNALT)	=	2 ounces of gold (20 hours of SNALT)

We have now reached the gold price of the commodity. Of course, for Marx this appearance hides the fact that it is the quantity of SNALT that renders commodities valuable.

Of course, in modern capitalist societies, the prices of commodities are stated in terms of units of paper money (e.g., so many dollars), not in terms of ounces of gold. For that reason, an explanation of paper money prices is required. During the nineteenth century, banks issued notes and governments issued paper bills in exchange for gold deposits. Depositors could deposit gold and then use the banknotes and paper bills to purchase commodities. When exchanges between commodities and paper occurred, the commodities were equated with the underlying money commodity, gold, rather than the paper itself. The paper simply served as a symbol or representative of the underlying money commodity. We can represent the paper expression of value as shown in Figure 4.7.

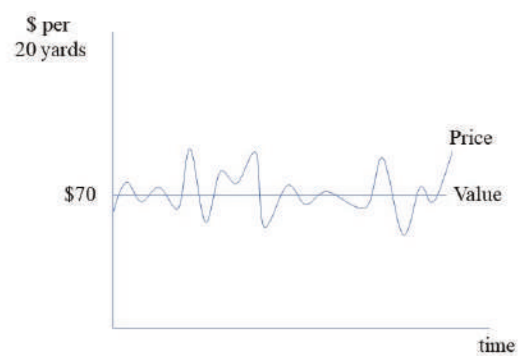
Figure 4.7: Paper as the Symbol of Value

20 yards of linen (20 hours of SNALT)	=	2 ounces of gold (20 hours of SNALT)	=	$70.00 (representative of 20 hours of SNALT)

Although the $70 requires virtually no SNALT for its production, it represents a claim to 2 ounces of gold in this example, which requires 20 hours of SNALT for its production. Therefore, the 20 yards of linen have a value of $70, and we have an explanation for the paper prices of commodities. Of course, the careful reader might object that paper money in modern capitalist economies is not directly convertible into gold. In fact, it is true that paper money today is **fiat money**, supported by nothing but the government's assurance that it has the legal title of money and the fact that people generally recognize its value. Although Marxian economists have an explanation for the value of fiat money that is consistent with Marx's theory of value, we will postpone this explanation until we reach the subject of monetary theory in Chapter 17.

The reader might ask at this stage what has happened to the laws of supply and demand. In Chapter 3, we learned of the important role these laws play in the determination of market prices within neoclassical economics. By explaining a commodity's value entirely in terms of the SNALT required for its production, was Marx denying the laws of supply and demand? Rather than denying these laws, Marx argues that these laws are secondary to the law of value. In other words, the laws of supply and demand can influence the **market price** of a commodity, but the value of the commodity depends on SNALT alone. For example, if the SNALT required for the production of the linen remains the same over time, the *value* of the linen will remain unchanged at $70, as shown in Figure 4.8.

Figure 4.8: Supply and Demand



Notice that the market *price* fluctuates considerably, however, in response to changes in the supply and demand for linen. The value of the linen represents the center of gravity around which the market price fluctuates and represents an average price over relatively long periods of time. It is here that we observe an important distinction between neoclassical and Marxian theory. Neoclassical economists only refer to market prices and the concept of value does not exist for them. Marxian economists recognize both concepts and consider the concept of value to be the more important object of analysis. As Marx wrote:

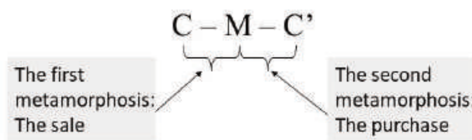
"[I]t is not the exchange of commodities which regulates the magnitude of their values, but rather the reverse, the magnitude of the value of commodities which regulates the proportion in which they exchange."[8]

Throughout this chapter, we will follow Marx's assumption throughout volume 1 of *Capital* in assuming that market prices are equal to values. That is, supply and demand do not force any significant deviations of price from value.

The Circulation of Commodities and the Meaning of Capital

In commodity exchange societies, commodities are constantly being bought and sold for money, which serves as the universal equivalent for the expression of value. Each sale and subsequent purchase can be represented as a **commodity circuit** through which the socially necessary labor of one commodity owner is equated with the socially necessary labor of another commodity owner. Indeed, Marx argued that all exchange within a commodity exchange society can be represented with a vast collection of interlocking commodity circuits. Marx chose to represent an individual commodity circuit symbolically as in Figure 4.9.

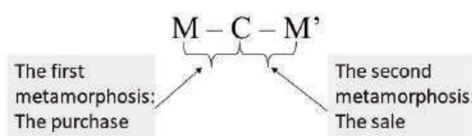
Figure 4.9: A Commodity Circuit



In Figure 4.9 a specific commodity (C) is transformed into its equivalent in money (M) when it is sold. The money is then used to buy a qualitatively different commodity (C'). The owner of the original commodity is able, in a sense, to transform the commodity into the new commodity through exchange. The hyphens connecting the symbols in the circuit indicate that a transformation is occurring. For example, the owner of five books might sell those books for $100. That money then purchases one desk. Throughout the process, the owner of the books equates the SNALT embodied in the books with the SNALT that the $100 represents and that the desk embodies. Let us suppose that each of these commodities represents 10 hours of SNALT. An exchange of equivalent values takes place then and the owner of books obtains something she wants in exchange (i.e., a desk) more than the thing she gives up (i.e., the books). It is a perfectly rational process for a commodity owner. This sphere of commodity circulation represents the realization of Adam Smith's Invisible Hand. Each commodity owner pursues her own self-interest by participating in these even swaps, and in the end, she obtains something that she desires more than the thing she loses.

Alongside these commodity circuits, Marx discovers another movement that exists within the capitalist mode of production. It is easy to miss, but Marx argues that this movement involves the purchase of commodities and their subsequent sale. Marx represents this movement symbolically as shown in Figure 4.10.

Figure 4.10: The General Formula for Capital



The problem that arises with this formula is that it appears to be a pointless exercise if equivalent values are exchanged. For example, if owners exchange equivalent values, then $100 representing 10 hours of SNALT might purchase a desk, which required 10 hours of SNALT for its production. If the owner then sells the desk for $100 representing 10 hours of SNALT, then owners have exchanged equivalent values but no expansion of the initial money has taken place. The owner of the initial money has no reason to risk the loss of her money by purchasing the desk only to sell it again. It makes more sense for her to keep her money out of circulation than to risk losing it. Notice that this problem does not arise in the case of a commodity circuit. Commodity owners exchange equivalent values and at the end of the circuit, the individual obtains a qualitatively different commodity from the one with which she began.

Because we have been assuming the exchange of equivalent values, the movement represented as M-C-M' makes little sense. That is, the starting point and the endpoint are *qualitatively* identical. The commodity owner begins with money and ends up with money. The only rational basis for engaging in such an exchange is to obtain a *quantitatively* different value (i.e., more money). Looking at Figure 4.10, the reader should note that the initial sum of money (M) purchases a commodity (C), which the commodity owner then sells for a different sum of money (M'). Because we will assume that M' is larger than M, we can write the following equation for M':

$$M' = M + \Delta M$$

That is, M' is larger than the original sum (M) by the amount ΔM. This increment (ΔM) is what Marx calls **surplus value**. When one uses money in this manner, Marx writes that money is transformed into **capital**. What then is capital? Capital in Marxian theory is money *and* it is commodities, but only when they take part in this specific movement. More generally, we can say that money is value in search of a greater value. Because capitalists are constantly striving to expand their capital in the endless search for surplus value, this movement never ceases in capitalist societies. The reader should notice how different this definition of capital is from the neoclassical definition of capital. For neoclassical economists, capital is only physical goods used to produce other goods. For Marxian economists, capital is the movement represented by M – C – M'.

Of course, if the person who advances the initial money capital ends up with a larger sum of money capital at the end of this process, then it might seem that we no longer can assume the exchange of equivalent values. For example, assume again that $100 representing 10 hours of SNALT buys a desk that requires 10 hours of SNALT for its production. This time, however, we will assume that when the desk is resold, it is sold for $120 representing 12 hours of SNALT. The surplus value in this case is $20, but the desk sells for more than it is worth. Commodity owners no longer exchange equivalent values.

It might seem that the exchange of unequal values must be the source of surplus value in capitalist societies. Marx, however, strongly rejects this explanation. According to Marx, if a seller of a commodity manages to sell a commodity at a price that exceeds its value (e.g., a $100 desk is sold for $120), then that individual seller manages to obtain a surplus value. The problem is that the buyer experiences a loss of value. What occurs then is a redistribution of value throughout society. The total value produced has not changed at all. As a result, Marx concludes that the answer to the question of where surplus value originates must lie outside the sphere of circulation.

To find his answer, Marx looks to the sphere of production to unlock the mystery of surplus value creation. Specifically, what if production makes it possible to enhance the value of the desk before the owner sells it? Once again, suppose that \$100 representing 10 hours of SNALT buys the desk, which requires 10 hours of SNALT for its production. If the value of the desk somehow increases from 10 hours of SNALT to 12 hours of SNALT during the production process, then the owner can sell it for its equivalent value of \$120. This example suggests that owners can exchange equivalent values in the sphere of circulation even as one owner realizes surplus value in exchange. The circulation of capital can now be represented in a slightly expanded form as follows:

$$M - C...C' - M'$$

In this expanded formula for capital, a change occurs to the original commodity (C) to transform it into a new commodity (C'). The question that we must ask is how this increase in the value of the commodity is possible.

The Commodity Labor-Power and the Determination of its Value

Marx explains that this expansion of the value of the commodity in the production process is possible because a very peculiar commodity is available in the marketplace that has a unique property. This commodity Marx calls **labor-power**. Labor-power is the capacity to perform work during a specific period (e.g., one hour, one day). Like any commodity, it has a use value and an exchange value. Even though it is not a tangible commodity, like a desk, a worker can sell it to a capitalist for a sum of money (i.e., a wage) and so it has exchange value. Its use value, however, is what is truly unique. It is unlike any other commodity that exists in a capitalist society because when it is used, it *creates* new value. Think about it for a moment. When a person buys a desk and consumes it over time, its value does not increase. If anything, when the person tries to sell it as a used item later, the price that it fetches will likely be below its initial value. When a capitalist buys labor-power and consumes it, however, she orders the worker to complete a specific type of work. As the worker performs this work, the worker performs socially necessary abstract labor. The SNALT performed adds value to the materials that the worker uses in production. It was this key insight that allowed Marx to explain the origin of surplus value.

Before we delve deeper into the details of how capitalists use labor-power to produce surplus value, we need to discuss the historical roots of this peculiar commodity. Labor-power is not a commodity that has always existed in human societies. In fact, it is specific to the capitalist mode of production. In ancient slave societies, for example, the enslaved people themselves were commodities. The slaves did not sell their ability to work for a specific period. Slaves possess no ownership rights themselves. Similarly, in feudal Europe, the serfs did not sell their labor-power to feudal lords in exchange for a wage. Serfs were bound to specific lands by tradition and custom. They handed a portion of their produce to the lords in exchange for protection and a place to live and work. It was only with the decline of feudalism and the spread of private ownership in land that landowners separated the serfs from the means of production and transformed them into wage laborers. For this reason, Marx identifies two key historical conditions that were required before labor-power became widely available as a commodity. First, workers needed to be *free* to enter into voluntary labor contracts and for that to happen, the legal system needed to recognize them as the rightful owners of their labor-power. Second, workers needed to be *forced to have this freedom* in the sense that if they did not sell their labor-power to capitalists, they would not be able to survive. In fact, as the serfs were forced from the land with the decline of feudalism in Europe, they had little choice but to migrate to the cities in search of wage work. Marx thus asserts that workers became **free in a double sense**.[9] Even though we take this socioeconomic relationship between workers and employers for granted as a natural and eternal relation, Marx insists that this socioeconomic relationship is historically specific. In Marx's words:

> "[N]ature does not produce on the one hand owners of money or commodities, and on the other hand men possessing nothing but their own labor-power."[10]

The widespread availability of labor-power is the defining feature of the capitalist mode of production. As we will see, capital cannot survive without it.

Because labor-power is a commodity with a use value and an exchange value, the quantity of SNALT required for its production must determine its value. It is clear what we mean when we state that a desk requires 10 hours of SNALT for its production, but what does it mean to state that one day's worth of labor-power requires 10 hours of SNALT for its production? Because labor-power is an intangible commodity, it is more difficult to imagine that it requires 10 hours of SNALT to produce it. What then determines the **value of labor-power**?

Marx explains that for a worker to perform work during the working day, he must have access to some essential commodities. For example, the worker requires food, clothing, housing, fuel, medical care, and possibly additional commodities each day. Each of these commodities requires SNALT for its production. Therefore, the value of the commodities that the worker requires each day to produce and reproduce her labor-power is what determines the value of labor-power. In other words, the value of labor-power depends on the value of the means of subsistence that are necessary to maintain the worker in her normal state as a working person.

Marx argues that the value of labor-power must also account for the dependents of the owner of labor-power because without such an allowance, future labor-power will not exist. He is also careful to point out that the value of labor-power contains an "historical and moral element," which means that the specific commodities that are regarded as the necessary means of subsistence are specific to time and place. In the United States in the twenty-first century, the necessary means of subsistence might include an automobile, cell phone service, and a health insurance policy. In the U.S. during the nineteenth century or in many parts of the developing world in the present day, none of these elements would be included.

Marx provides a precise formula for the value of one day's labor-power as follows:[11]

$$The\ Value\ of\ One\ Day's\ Labor-Power = \frac{365A+52B+4C+...}{365}$$

In this formula, A represents the value of the commodities that the worker requires daily, B represents the value of the commodities that the worker requires weekly, and C represents the value of the commodities that the worker requires quarterly. We multiply each value by the number of periods within a year. The numerator, therefore, represents the **annual value of labor-power**. If we divide the annual value of labor-power by the number of days in the working year (assumed to be 365 days), then we obtain the **daily value of labor-power**. We could further divide this amount by the number of working hours in the workday to obtain the **hourly value of labor-power**.

For example, suppose that A represents the value of food required daily, B represents the value of fuel required weekly, and C represents the value of the rent for housing required quarterly. We may express these values in terms of SNALT or in dollar terms. If we assume that A equals \$10 of food per day, B equals \$73 of fuel per week, and C equals \$3,650 of housing per quarter, then we have the following results assuming an 8-hour workday:

$$Annual\ Value\ of\ Labor-Power = 365(10)+52(73)+4(3650) = \$22,046\ per\ year$$

$$Daily\ Value\ of\ Labor-Power = \frac{\$22{,}046\ per\ year}{365\ days\ per\ year} = \$60.40\ per\ day$$

$$Hourly\ Value\ of\ Labor-Power = \frac{\$60.40\ per\ day}{8\ hours\ per\ day} = \$7.55\ per\ hour$$

We will continue to assume that prices equal values at this stage. That is, we assume that the price of labor-power equals the value of labor-power. In other words, we are ruling out the possibility at this stage that fluctuations in the supply and demand for labor-power might cause the price of labor-power to rise above or fall below its value.

The Length of the Working Day and the Production of Surplus Value

Before we can unlock the secret of surplus value production, we must introduce a few more concepts. Labor-power has taken its place within the world of commodities that are available in the marketplace. Despite its peculiar characteristic, its value is determined like that of all the other commodities. Figure 4.11 shows one day's labor-power as a single commodity among many other commodities.

Figure 4.11: The Inclusion of Labor-Power in the World of Commodities

$$\left.\begin{array}{c}\text{20 yards of linen}\\ \text{1 coat}\\ \text{10 lbs. of tea}\\ \text{½ ton of iron}\\ \text{1 day of labor-power}\\ \text{etc.}\\ \text{(each requires 6}\\ \text{hours of SNALT)}\end{array}\right\} = \begin{array}{c}\text{\$90.00}\\ \text{(represents}\\ \text{6 hours of SNALT)}\end{array}$$

In this example, we assume that 6 hours of SNALT are required to produce and reproduce a day's worth of labor-power. It sells for \$90, which represents an equivalent amount of SNALT. Therefore, if labor-power sells at its full value of \$90, then a capitalist fully compensates a worker for the commodity she sells. This point is important. If we make this assumption, we cannot draw the conclusion that the worker is exploited in the realm of commodity circulation. An additional important assumption is that \$90 represents 6 hours of SNALT. This ratio allows us to represent any amount of SNALT in monetary terms. In the equation below, we simplify this ratio, which Marxian economists refer to as the **monetary expression of labor time (MELT)**.

$$MELT = \frac{\$90\ per\ day}{6\ hours\ per\ day} = \$15\ per\ hour\ (of\ SNALT)$$

The reader should be aware that the MELT is not the same as the hourly value of labor-power (i.e., the wage). The MELT only tells us that \$15 represents 1 hour of SNALT because the amount of gold that \$15 represents (under the gold standard) requires 1 hour of SNALT for its production.

We only have a few more concepts to introduce and then we can bring together our concepts in an example to reveal how capitalist production produces surplus value. When a capitalist advances a specific amount of money capital (M), the capitalist advances a portion of it for the purchase of the means of production. The means of production in a specific production process include the **instruments of labor** (e.g., tools, machinery, and specialized equipment) and the **objects of labor** (e.g., raw materials, ingredients, component parts). Neoclassical economists refer to all these means of production simply as capital. The portion of the money capital that capitalists advance for the purchase of means of production, Marxian economists refer to as **constant capital (c)**. The other part of the money capital that is

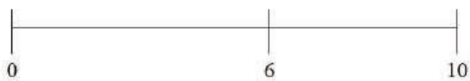
advanced purchases labor-power. This part of the capital, Marxian economists refer to as **variable capital (v)**. The reason for these labels will become clearer after we have worked through an example in which capitalist production produces surplus value. The money capital (M) may then be expressed as follows:
$M = c + v$

As in neoclassical theory, we will make certain assumptions, and then analyze the situation given the restrictions that we have defined. We will assume that a capitalist hires one worker to work a 10-hour workday. The capitalist advances \$300 of constant capital for means of production and \$90 of variable capital for labor-power. That is:
$M = c + v = 300 + 90 = \$390$

We will also assume that the MELT is \$15 per hour of SNALT as above and that the worker will produce a total product (TP) of 225 lbs. of sugar during the workday. With this information, we can uncover the origin of surplus value.

Figure 4.12 shows a timeline representing the length of the working day for this example.

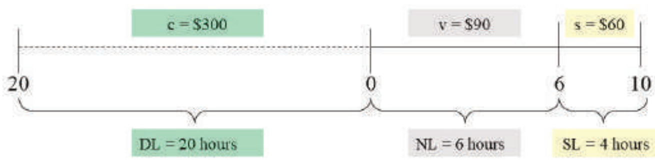
Figure 4.12: The Division of the Working Day



As each hour passes, the worker creates more new value for the capitalist in the form of commodities (i.e., finished sugar). By the end of the sixth hour, the worker has created enough new value to equal the variable capital advanced. Since $15 is the monetary equivalent of an hour of SNALT, $90 (= $15 per hour times 6 hours) is the new value created in these 6 hours. The key point to notice, however, is that the worker has agreed to work for the entire 10 hours. As a result, the worker performs 4 hours of additional labor. Surplus value is possible because of the performance of this extra labor.

Of course, the 10 hours of SNALT that the worker performs during the workday to produce the finished sugar is not the only labor time embodied in the product. The labor process transfers the SNALT embodied in the means of production to the final product as well. Figure 4.13 breaks down the total value of the sugar produced in the day into its constant capital, variable capital, and surplus value components.

Figure 4.13: The Component Parts of the Total Value



Each of these monetary amounts corresponds to an equivalent value in terms of labor time. For example, the SNALT embodied in the means of production is referred to as **dead labor (DL)**. The SNALT required to produce a value equivalent to the variable capital we call **necessary labor (NL)**. We call this labor necessary labor because it is the amount of SNALT necessary to produce a value equivalent to the required means of subsistence, and we should not confuse it with the concept of socially necessary labor.[12] Finally, the SNALT that the worker performs over and above the necessary labor, we refer to as **surplus labor (SL)**. The entire amount of labor that the worker performs during the workday is called **living labor (LL)**. It follows that:

$$LL = NL + SL$$

Because the MELT makes it possible to convert between monetary values and labor time equivalents, we can work through this problem as follows:

Determine the amount of necessary labor time.

- $NL = \frac{v}{MELT} = \frac{\$90}{\frac{\$15}{hour}} = 6\ hours$

Determine the amount of surplus labor time.

- $SL = LL - NL = 10\ hours - 6\ hours = 4\ hours$

Determine the amount of surplus value (s) produced.

- $s = SL \cdot MELT = 4\ hours \cdot \frac{\$15}{hour} = \$60$

Determine the amount of dead labor time embodied in the means of production.

- $DL = \frac{c}{MELT} = \frac{\$300}{\frac{\$15}{hour}} = 20\ hours$

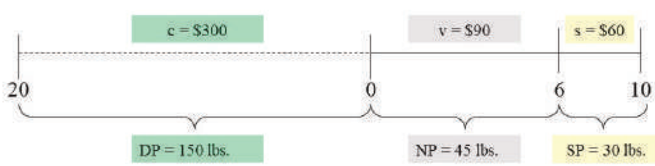
We can now see that the **total value (TV)** of the sugar produced in the day and the **total labor (TL)** embodied in the final product for the day are calculated as follows:
$TV = c + v + s = 300 + 90 + 60 = \$450$
$TL = DL + NL + SL = 20 + 6 + 4 = 30\ hours$

It should be noted that \$450/30 hours equals \$15 per hour, which is also the MELT.

We can also calculate the monetary value or price (p) of one pound of sugar by dividing the total value of the sugar by the total product as follows:
$p = \frac{TV}{TP} = \frac{\$450}{225\ lbs.} = \$2\ per\ lb.$

Just like the constant capital, the variable capital, and the surplus value correspond to the dead labor, necessary labor, and surplus labor, respectively, they also correspond to the specific parts of the total product. That is, the constant capital corresponds to what we can call the **dead product (DP)**, the variable capital corresponds to the **necessary product (NP)**, and the surplus value corresponds to the **surplus product (SP)**, as shown in Figure 4.14.

Figure 4.14: The Relationship between the Value Components and the Total Product



To calculate each of the components of the total product, it is only necessary to divide each of the monetary magnitudes by the price of sugar as follows:

$$DP = \frac{c}{p} = \frac{\$300}{\$2\ per\ lb.} = 150\ lbs.$$

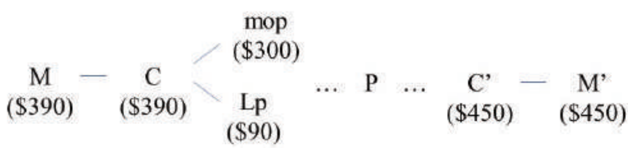
$$NP = \frac{v}{p} = \frac{\$90}{\$2\ per\ lb.} = 45\ lbs.$$

$$SP = \frac{s}{p} = \frac{\$60}{\$2\ per\ lb.} = 30\ lbs.$$

The total product should be simply the sum of these component parts. That is:
$TP = DP + NP + SP = 150 + 45 + 30 = 225\ lbs.$

We now have a complete picture of the process by which surplus value is created. We can also expand the circuit of capital to gain a clearer picture of the entire process as shown in Figure 4.15.

Figure 4.15: The Expanded Version of the Circuit of Capital



This expanded version of the circuit of capital shows that the commodities (C) that are purchased using the initial money capital at the beginning of the circuit include means of production (mop) and labor-power (Lp). The circulation process is then interrupted with the production phase (P). Once production is complete, a new commodity (C') emerges that contains surplus value. It then sells for an amount of money (M'), and the capitalist realizes the surplus value. The capital has expanded. Marx's theory suggests that surplus value exists because the capitalist exploits the worker in the production phase of the circuit. Even though the worker is responsible for the entire new value created, the capitalist owns the final product and decides how the proceeds from its sale are used. Therefore, **exploitation** within the capitalist mode of production refers to the production of surplus value by workers and its appropriation and distribution by capitalists. It is here that we see why Marx's theory is an impressive blend of positive and normative analysis. The reference to this process as one involving exploitation indicates Marx's condemnation of it. At the same time, it has a technical meaning that describes how a worker produces more value then she receives in the form of a wage.

The reader should be able to understand more clearly now why the labels constant capital and variable capital have been used. The reason for this distinction is that labor-power has the peculiar property that it creates new value when consumed. It alone is responsible for the expansion of the capital value, and so Marxian economists regard the capital advanced for the purchase of labor-power as variable in nature. By contrast, the value of the means of production is only *preserved and transferred* to the final product. As a result, their value does not expand and so this part of the capital Marxian economists regard as constant in nature.

The reader might object to this entire analysis because capitalists do not pay workers one lump sum for the day but rather pay workers for each hour worked. It should be noted, however, that in this case, the hourly wage would be $9 per hour (= $90 per day/10 hours per day) so even though it looks like the worker is compensated for every hour she works, this appearance is a false one.

The worker works only part of the hour for herself and the other part for the capitalist. Remember, each hour worked creates the value equivalent of $15, according to the MELT. Hence, $9 of value created in an hour corresponds to necessary labor whereas the other $6 of value created in an hour corresponds to surplus labor when we view the problem in this manner. Hourly wages rather than daily wages do nothing to change the fact that the capitalist appropriates a part of the value that the worker produces.

The Rate of Surplus Value and the Rate of Profit

Marx also introduces two important measures that correspond to this analysis. The first measure he calls the **rate of surplus value** or the **degree of exploitation** of labor-power. This measure is of great importance to the worker because it measures the extent to which the capitalist exploits the worker in the production process. Marx defines it as follows:

$$Rate\ of\ surplus\ value = \frac{s}{v} = \frac{SL}{NL} = \frac{\$60}{\$90} = \frac{4\ hours}{6\ hours} = 66.67\%$$

The rate of surplus value measures the degree to which the worker works for the capitalist rather than for herself. The higher is the rate of surplus value, the higher is the degree of exploitation.

A second measure of importance is the **rate of profit**. The rate of profit measures the rate at which the capital value has expanded during the production process. It is of greatest interest to the capitalist because it is how the capitalist measures the profitability of her investment. Marx defines it as follows:

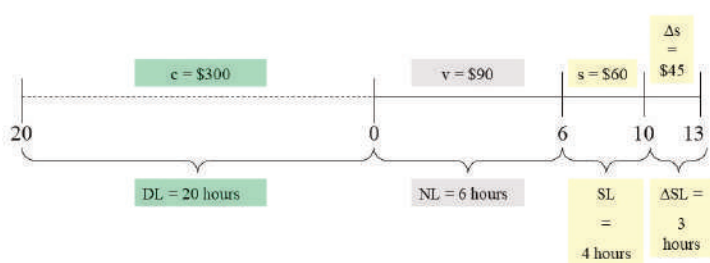
$$Rate\ of\ profit = \frac{s}{c+v} = \frac{\$60}{\$390} = 15.38\%$$

The higher is the rate of profit, the more profitable is the investment. That is, a higher rate of profit indicates a more rapid expansion of the capital value.

The Production of Absolute Surplus Value and Relative Surplus Value

Because the capitalist is interested in extracting as much surplus value as possible from the worker, it is important to consider which measures the capitalist can take to increase the amount of surplus value produced. One way to increase the amount of surplus value produced is through an extension of the length of the working day. Marx refers to an increase in the production of surplus value due to an extension of the workday as an increase in **absolute surplus value**. Figure 4.16 shows the modifications to our example that stem from a lengthening of the workday.

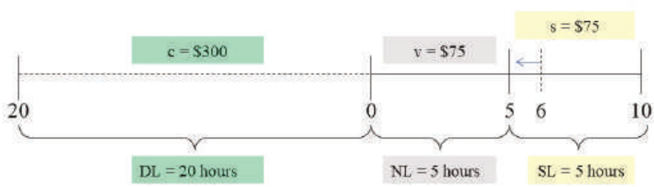
Figure 4.16: The Production of Absolute Surplus Value



In this example, the working day has been extended by 3 hours. As a result, the worker performs 3 hours of additional surplus labor (ΔSL). Multiplying this amount of additional surplus labor by the MELT, we obtain the additional surplus value (Δs) produced of $45. It should be noted that because the constant capital and the variable capital have not changed, this increase in the surplus value will increase both the rate of surplus value and the rate of profit. The rate of surplus value rises to 116.67% (= $105/$90), and the rate of profit rises to 26.92% (= $105/$390). That is, the worker is exploited to a greater degree, and the capital expands more quickly as a result.

The second way in which the amount of surplus value may increase is through a new division of the working day due to a change in the value of labor-power. For example, suppose that productivity increases in the sectors that produce the means of subsistence for the worker. In that case, the values of these necessities will decline because less SNALT is required for their production. It will now be cheaper for the worker to purchase those commodities deemed necessary to keep her in a condition that is normal for a working person and that are required for her to perform her work at an average level. The value of labor-power, therefore, declines. As a result, the capitalist will not need to advance as much variable capital as previously. Because the necessary labor decreases with this reduction in the variable capital advanced, a larger portion of the workday is left over for the performance of surplus labor. Marx refers to an increase in the surplus value produced arising from a new division of the working day as an increase in the production of **relative surplus value**. Figure 4.17 shows how the original example is modified due to a cheapening of the value of labor-power.

Figure 4.17: The Production of Relative Surplus Value



In this case, the variable capital has declined from \$90 per day to \$75 per day due to the reduction in the value of labor-power. The necessary labor time falls to only 5 hours (= \$75/\$15 per hour). Due to the reduction in the necessary labor time to 5 hours and the fact that the workday is fixed at 10 hours, the surplus labor time performed rises from 4 hours to 5 hours. The surplus value also rises then to \$75. In this example, the rate of surplus value rises to 100% (= \$75/\$75) for the twofold reason that more surplus value is produced and less variable capital is advanced. In addition, the rate of profit increases in this example to 20% (= \$75/\$375). The increase in the rate of profit is due to the increase in surplus value, but it is also due to the reduction in the amount of capital that must be advanced.

Of course, changes do not always work in the favor of capitalists. If workers form a union and demand shorter working hours, then the rates of surplus value and profit will decline. In addition, if the means of subsistence rise in value, then the capitalist will appropriate less surplus value because she must pay the worker a higher wage. As a result, the rates of surplus value and profit will fall. In this case, the worker is no better off despite the higher wage, however, because the higher wage is paid solely to allow the worker to buy the same quantity of the means of subsistence as before.

As we conclude this chapter, we should place special emphasis on the uniqueness of the capitalist process of production. In all human societies, the production of use values is essential. Indeed, in Marxian economics, the **labor process** refers to any purposeful activity aimed at the production of use values. In addition to purposeful productive activity, Marx explains that the labor process includes the object of labor and the instruments of labor. The labor process is an important element within the capitalist mode of production as well, but within capitalism it overlaps with another process that Marx calls the **valorization process**. The valorization process refers to the process by which surplus value is created. As we have seen, the capitalist owns labor-power, the means of production, and the commodity that is ultimately produced. The goal of the capitalist then is not to produce use values but rather commodities containing surplus value. As this chapter has demonstrated, the origin of surplus value rests in the productive consumption of labor-power, which leads to the creation of new value that exceeds its own value. It is this valorization process that gives the capitalist production process its uniqueness and that is the defining characteristic of the capitalist mode of production.

Following the Economic News[13]

Foxconn is a Taiwanese electronics giant that produces iPhones, iPads, and electronics for several major producers of electronic devices, including Apple, Sony, and Nokia. Three Foxconn workers committed suicide at one of the firm's production plants in the central city of Zhengzhou, China during a three-week period in the spring of 2013. In 2010 at least 13 Foxconn workers committed suicide at the firm's production plants in China. The cause appears to be harsh working conditions, long working hours, and low pay. A 2011 report into conditions at Foxconn's Chinese plants concluded that workers were being exposed to harmful disease, were being humiliated by management, and were not being paid enough to meet their basic needs. According to the report, workers at Foxconn's Chengdu factory only earn $186 per month. One Chinese Foxconn worker explains that he stands for at least 14 hours each day producing cases for the iPad.

In terms of the Marxian theory developed in this chapter, it sounds like Foxconn has been maximizing the length of the working day to extract as much absolute surplus value as possible from its workers. It also sounds like the firm has been paying wages that are below the value of labor-power because the workers are not able to earn enough to purchase necessities. To the extent that the firm has been able to reduce wages in this manner, it has been able to extract a larger amount of relative surplus value. In this situation, the workers are made worse-off due to the decline in wages because necessities have not become any cheaper. To the extent that it is possible for the firm to push the price of labor-power below its value, this tactic is another way to increase the amount of surplus value produced and thus the rate of profit. In response to the public outcry, Foxconn has promised to raise pay rates, but the report found that the firm has been cancelling the pay increases by refusing to provide a separate food and housing allowance. As the recent suicides indicate, however, a firm's ability to force the price of labor-power below its value has limits, and corrective action will need to be taken to ensure that a productive workforce exists in the future. As a temporary measure, the firm has installed nets to catch workers jumping from rooftops and has required workers to sign "no-suicide" pacts.

Summary of Key Points

1. Marx's theory of capitalism represents a merger of Hegelian philosophy, French political theory, and British political economy.
2. The Marxian circular flow model is different from the neoclassical circular flow model in that the Marxian model emphasizes a social class hierarchy, persistent unemployment, inequality in consumption, and the distinctive nature of work as a factor contributed to industry.
3. For a thing to be considered a commodity, it must be both a use value and an exchange value.
4. The value of a commodity depends directly on the amount of socially necessary abstract labor time (SNALT) required for its production.
5. Commodities exchange for units of paper money in Marxian economics because the units of paper represent definite quantities of gold that require

equivalent amounts of SNALT for their production.

6. Commodity circuits are represented as C-M-C' whereas the circuit of capital is represented as M-C-M'.
7. Labor-power is a special commodity that can create more value than it is worth when it is consumed.
8. The value of labor-power depends on the SNALT required for its production and reproduction (i.e., the SNALT embodied in the means of subsistence).
9. The monetary expression of labor time (MELT) makes possible the conversion of any amount of SNALT into its equivalent in money.
10. The exploitation of labor-power may be increased either through an extension of the working day or through a cheapening of the value of labor-power.

List of Key Terms

Dialectical idealism

Dialectical materialism

Bourgeoisie

Proletariat

Marxian circular flow model

Reserve army of the unemployed

Necessities

Luxuries

Net social product

Wealth

Commodity

Use value

Exchange value

Concrete, private labor

Abstract, social labor

Value

Socially necessary abstract labor time (SNALT)

Law of value

Simple form of value

Expanded form of value

Universal equivalent

General form of value

Money form of value

Fiat money

Market price

Commodity circuit

Surplus value

Capital

Labor-power

Free in a double sense

Value of labor-power

Annual value of labor-power

Daily value of labor-power

Hourly value of labor-power

Monetary expression of labor time (MELT)

Instruments of labor

Objects of labor

Constant capital (c)

Variable capital (v)

Dead labor (DL)

Necessary labor (NL)

Surplus labor (SL)

Living labor (LL)

Total value (TV)

Total labor (TL)

Dead product (DP)

Necessary product (NP)

Surplus product (SP)

Exploitation

Rate of surplus value (or degree of exploitation)

Rate of profit

Absolute surplus value

Relative surplus value

Labor process

Valorization process

Problems for Review

Assume the following initial conditions for problems 1-7. Be sure to include the proper units in each of your answers.

- The working day is fixed at 11 hours.
- The capitalist hires one worker for the day for which the worker is paid $63.
- The capitalist buys the means of production (tools, raw materials) for $189.
- The worker produces 950 lbs. of sugar during the working day.
- The worker consumes (i.e., uses up) all the means of production.
- The monetary expression of labor time is $7 per 1 hour of SNALT. That is, it takes one hour to produce the gold that the $7 represents under the gold standard.

Complete the following problems:

1. Draw a timeline representing the length of the working day in this example.
2. Identify the necessary labor, the surplus labor, and the dead labor.
3. Identify the constant capital, the variable capital, and the surplus value.
4. What is the total (monetary) value of the day's product? What is the (monetary) value (or price) of each pound of sugar?
5. What is the rate of surplus value? What is the rate of profit?
6. If the working day increases to 12 hours, calculate the new rate of surplus value and the new rate of profit? What happened to each (increased, decreased, constant)?
7. Returning to the initial conditions when the working day is set at 11 hours (that is, disregarding the change that occurs in problem 6), calculate the new rate of surplus value and the new rate of profit when the wage paid to the worker rises to $70 for the day. What happened to each (increased, decreased, constant)?
8. If a worker requires $12 worth of food per day, $28 worth of clothing per week, and $985 of rent per month, calculate the following:

- The annual value of labor power
- The daily value of labor power (assuming a 365-day working year)
- The hourly value of labor power (assuming an 8-hour working day)

Notes

1. For a helpful overview of the way in which Marx’s thought represents a synthesis of three major bodies of European thought, see Rosser and Rosser (2004), p. 56-59.
2. Ibid. p. 56.
3. Ibid. p. 57.
4. Ibid. p. 58.
5. Figure 4.1 is a slightly modified version of Prof. Nell's original figure.
6. Marx reasoned that because the concept of use value is entirely qualitative in nature, it cannot explain the quantitative relationship of a commodity to other commodities (i.e., exchange value).
7. See Marx (1990), p. 131.
8. Marx (1990), p. 156.
9. Marx (1990), p. 272.
10. Ibid. p. 273.
11. Marx (1990), p. 276.
12. Marx offers a cautionary note to this effect. See Marx (1990), p. 325, footnote 5.
13. Moses, Asher. “Meet the workers dying to meet your iPad 2 demand.” *The Sydney Morning Herald*. May 9, 2011; and “Three suicides at Apple supplier’s China factory.” *The Sydney Morning Herald*. May 21, 2013.

PART III

PART TWO: PRINCIPLES OF MICROECONOMIC THEORY

Chapter 5

THE NEOCLASSICAL CONCEPT OF ELASTICITY

Goals and Objectives:

In this chapter, we will do the following:

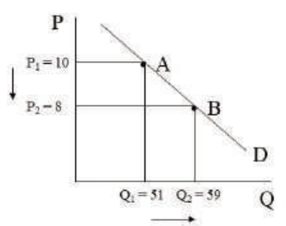
1. *Define* the neoclassical concept of price elasticity of demand and explain its importance
2. *Calculate* and *interpret* the price elasticity of demand
3. *Learn* how and when to use the arc elasticity formula and the point elasticity formula
4. *Identify* key determinants of the price elasticity of demand
5. *Explain* the relationship between the price elasticity of demand and sales revenue
6. *Define*, *calculate*, and *interpret* other key measures of elasticity

The Need for a Measure of Consumer Responsiveness

In Chapter 3, we spent a great deal of time discussing the law of demand, which states that, other things equal, as the price of a good declines, the quantity demanded rises, and vice versa. In this chapter, we are interested in more than the *direction* of the change in quantity demanded when the price changes. That is, we wish to know the *magnitude* of the change in quantity demanded when the price changes. As a result, we require a measure of the change in quantity demanded when the price changes by some amount. This measure of the responsiveness of the consumer to a change in price is known to neoclassical economists as the **price elasticity of demand.**

The reader may wonder why neoclassical economists require this special measure of consumer responsiveness to a price change. After all, the slope of the demand curve tells us the change in quantity demanded for a specific change in the price of the good. Figure 5.1 illustrates how to calculate the slope of a linear demand curve.

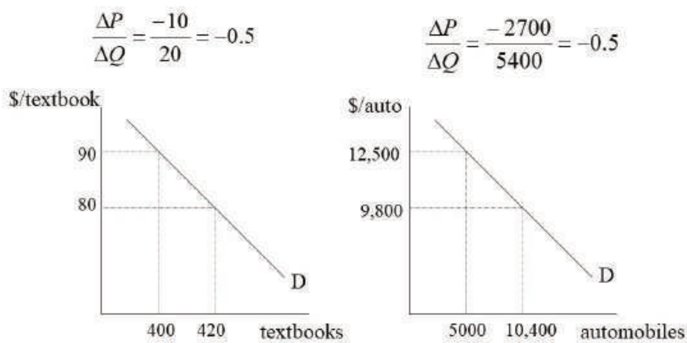
Figure 5.1: Calculating the Slope of a Linear Demand Curve



$$\frac{\Delta P}{\Delta Q}=\frac{P_2-P_1}{Q_2-Q_1}=\frac{8-10}{59-51}=\frac{-2}{8}=-\frac{1}{4}$$

In Figure 5.1, the price of the good declines from \$10 per unit to \$8 per unit and the quantity demanded rises from 51 units to 59 units of the good. The movement from point A to point B reveals that the slope is equal to -1/4. The slope appears to provide us with a measure of the responsiveness of the consumer to a price change. A serious problem exists, however, with this measure of consumer responsiveness. Specifically, it is not a *unit-free* measure of consumer responsiveness, and therefore, it does not permit comparisons of consumer responsiveness across different markets. To better grasp this point, consider Figure 5.2.

Figure 5.2: The Inadequacy of Slope as a Measure of Consumer Responsiveness



In Figure 5.2, the slopes of the two demand curves are both equal to -1/2, suggesting that the consumers are equally responsive to price changes in the two markets. The problem, however, is that we are comparing apples and oranges, or in this case, textbooks and automobiles. In other words, the slope of the demand curve is not a unit-free measure of consumer responsiveness. More generally, we might ask in which units the slope is measured. The price of a good is measured in dollars per unit of the good ($/Q). The slope then, which is calculated as ΔP/ΔQ, must be measured in terms of ($/Q)/Q or $/Q^2.[1] Clearly, this measure is not unit-free and will not allow direct comparisons of consumer responsiveness across markets.

Calculating and Interpreting the Price Elasticity of Demand

The price elasticity of demand is a measure that avoids the shortcomings of the slope of the demand curve as a measure of consumer responsiveness. It is calculated in the following way, where E_D refers to the price elasticity of demand:

$$E_D = \frac{\%\Delta Q^D}{\%\Delta P}$$

For example, when the price of a good rises by 6% and the quantity demanded of the good falls by 12%, then the price elasticity of demand is equal to -2 as shown in the following calculation:

$$E_D = \frac{\%\Delta Q^D}{\%\Delta P} = \frac{-12}{6} = -2$$

The careful reader will notice two important differences between this measure and the slope of the demand curve. First, the price elasticity of demand places Q in the numerator and P in the denominator, which suggests that this measure will be inversely related to slope. Second, the price elasticity of demand includes *percentage* changes as opposed to *absolute* changes in P and Q.

The next question we must ask, however, is whether this measure of consumer responsiveness is unit-free. Only then will it be possible to make comparisons of consumer responsiveness across markets. It turns out that it is unit-free. Consider, for example, a child that begins the month weighing 100 lbs. At the end of the month, the child weighs 105 lbs. Because the child's original weight (W) is 100 lbs. and the change in the child's weight (ΔW) is 5 lbs., the percentage change in the child's weight (ΔW/W) is 5 lbs./100 lbs. or 5%. The reader can easily see that the units cancel out in this calculation, leaving us with a pure or unit-free number. It is the same way with all calculations involving percentages. In the case of the price elasticity of demand, we can rewrite the formula in the following way:

$$E_D = \frac{\%\Delta Q^D}{\%\Delta P} = \frac{\frac{\Delta Q}{Q}}{\frac{\Delta P}{P}}$$

The reader should note that cancellation of the units in the numerator and denominator of this fraction leads to the calculation of a unit-free measure of consumer responsiveness.

If we return to the example in which the price elasticity of demand was calculated to be -2, it is natural to wonder why the result has included a negative sign. This result should come as no surprise because the law of demand states that price and quantity demanded are inversely related, other things held constant. In fact, we should almost always expect a negative sign when we calculate the price elasticity of demand. For this reason, neoclassical economists typically omit the negative sign when referring to the price elasticity of demand and instead refer to its absolute value. A neoclassical economist would state then that the price elasticity of demand is 2 even though she really means -2. Our formula for the price elasticity of demand can be modified as follows, and the previous calculation would be carried out as shown:

$|E_D| = |\frac{\%\Delta Q^D}{\%\Delta P}| = |\frac{-12\%}{6\%}| = |-2| = 2$

The absolute value of the price elasticity of demand is used then as a kind of shorthand. It is useful as well for another reason. When the absolute value of the price elasticity of demand is equal to 2, it can be interpreted to mean that a 1% rise in price causes a 2% reduction in quantity demanded, other things equal. If the absolute value is equal to 3, then a 1% rise in the price causes a 3% reduction in quantity demanded. In the latter case, we observe a larger response from consumers given a 1% rise in price. When consumers are more responsive to a price change, it is said that demand is more **elastic**. Because we are referring to the absolute value, the larger the number is, the more elastic the demand will be. If we were using the negative values, then we would be forced into the less natural position of claiming that the lower the value, the greater the elasticity of demand.

More generally, the price elasticity of demand may take on any value from zero to positive infinity (in absolute value terms). Depending on the specific value, different language is used to refer to the elasticity of demand. Below are the phrases used for each possible range of values. The symbol "⟺" indicates that the two statements are logically equivalent.

$|E_D| > 1 \Longleftrightarrow Demand\ is\ elastic$

$|E_D| = 1 \Longleftrightarrow Demand\ is\ unit\ elastic$

$|E_D| < 1 \Longleftrightarrow Demand\ is\ inelastic$

$|E_D| = 0 \Longleftrightarrow Demand\ is\ perfectly\ inelastic$

$|E_D| = \infty \Longleftrightarrow Demand\ is\ perfectly\ elastic$

It has already been emphasized that a larger price elasticity of demand implies greater responsiveness on the part of consumers. Nevertheless, more must be said to explain why a value of 1 is so important to the interpretation of this value. To understand the reason, we return to the definition of the price elasticity of demand and observe the following:

$|\frac{\%\Delta Q_D}{\%\Delta P}| = 1 \Rightarrow |\%\Delta Q_D| = |\%\Delta P| \Rightarrow Demand\ is\ unit\ elastic$

That is, when demand is **unit elastic**, a 5% rise in price leads to a 5% reduction in quantity demanded. The consumers, in other words, respond by an equal percentage amount to a price change. For this reason, the value of 1 is important for the purposes of interpreting the price elasticity of demand.

When the price elasticity of demand exceeds 1, we have the following:

$|\frac{\%\Delta Q_D}{\%\Delta P}| > 1 \Rightarrow |\%\Delta Q_D| > |\%\Delta P| \Rightarrow Demand\ is\ elastic$

This result indicates that demand is elastic whenever the consumers respond by a greater percentage amount than the price change (in absolute value terms).

When the price elasticity of demand is less than one, we have the following:

$|\frac{\%\Delta Q_D}{\%\Delta P}| < 1 \Rightarrow |\%\Delta Q_D| < |\%\Delta P| \Rightarrow Demand\ is\ inelastic$

This result indicates that demand is **inelastic** whenever the consumers respond by a smaller percentage amount than the price change (in absolute value terms).

When the price elasticity of demand is equal to 0, we have the following:

$|\frac{\%\Delta Q_D}{\%\Delta P}| = 0 \Rightarrow |\%\Delta Q_D| = 0 \Rightarrow Demand\ is\ perfectly\ inelastic$

This result indicates that demand is **perfectly inelastic** whenever the consumers do not respond at all to a given percentage change in the price of the good.

Finally, when the price elasticity of demand is equal to ∞, we have the following:
$|\frac{\%\Delta Q_D}{\%\Delta P}| = \infty \Rightarrow |\%\Delta P| = 0 \Rightarrow Demand\ is\ perfectly\ elastic$

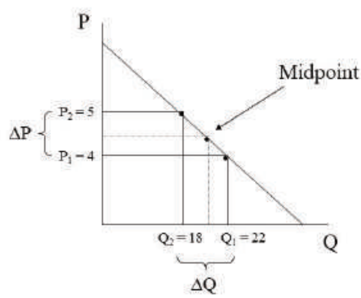
Division by zero causes the fraction to be undefined. In this case, however, we can imagine a very tiny percentage price change such that it approaches 0 and the entire fraction, therefore, approaches infinite. In other words, even for the smallest (nearly zero) change in price, we obtain an opposite change in quantity demanded. That is consumers are perfectly responsive to any price change at all, and demand is said to be **perfectly elastic**.

Notice that with this measure we can now compare the responsiveness of orange consumers and apple consumers if we know the price elasticity of demand for each. If the price elasticity of demand for oranges is 2 and the price elasticity of demand for apples is 3, then we conclude that apple consumers are relatively more responsive than orange consumers to price changes. In a sense, we are free to compare apples and oranges, which was not possible when we only knew the slopes of the demand curves.

The Arc Elasticity and Point Elasticity Formulas

When we know the percentage changes in price and quantity demanded, it is very easy to calculate the price elasticity of demand. We simply use the definition of the price elasticity of demand that was provided in the previous section. At times, however, we might find ourselves only knowing the absolute changes in price and quantity demanded. Without direct knowledge of the percentage changes in price and quantity demanded, we must find some way of obtaining this information from the information we do have. For example, consider the graph in Figure 5.3.

Figure 5.3: A Situation that Requires the Arc Elasticity Formula



In this case, we know that the price of the good rises from \$4 per unit to \$5 per unit and that the quantity demanded subsequently falls from 22 units to 18 units. That is, we know only the absolute changes in price and quantity demanded. To convert our formula for the price elasticity of demand from one that uses *percentage* changes to one that uses *absolute* changes, we need to return to our initial definition and make the conversion as follows:

$$E_D = \frac{\%\Delta Q^D}{\%\Delta P} = \frac{\frac{\Delta Q}{Q}}{\frac{\Delta P}{P}} = \frac{\frac{Q_2 - Q_1}{Q}}{\frac{P_2 - P1}{P}}$$

This formula appears to solve our problem because we can now use absolute quantities and prices to carry out a calculation of the price elasticity of demand. Unfortunately, the formula has one major defect. The problem with this formula is that it is not immediately clear which Q and P should be used in the denominator of each fraction. When neoclassical economists first considered this problem, they decided that it would be completely arbitrary to use Q_1 and P_1 as opposed to Q_2 and P_2. The issue is important because the calculation is different depending on the choice that is made. As a result, they concluded that it made the most sense to use the average quantity demanded and the average price when carrying out the calculation. As a result, the formula changes to the following:

$$E_D = \frac{\%\Delta Q^D}{\%\Delta P} = \frac{\frac{\Delta Q}{Q}}{\frac{\Delta P}{P}} = \frac{\frac{Q_2-Q_1}{Q}}{\frac{P_2-P1}{P}} = \frac{\frac{Q_2-Q_1}{\frac{Q_1+Q_2}{2}}}{\frac{P_2-P1}{\frac{P_1+P_2}{2}}}$$

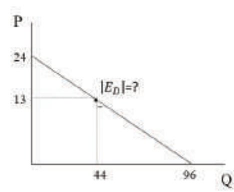
If we write the **arc elasticity** formula in terms of its absolute value and carry out the calculation using the information provided in Figure 5.3, then we obtain the following:

$$|E_D| = |\frac{\frac{Q_2-Q_1}{\frac{Q_1+Q_2}{2}}}{\frac{P_2-P1}{\frac{P_1+P_2}{2}}}| = |\frac{\frac{18-22}{\frac{22+18}{2}}}{\frac{5-4}{\frac{4+5}{2}}}| = |\frac{\frac{-4}{20}}{\frac{1}{9/2}}| = |\frac{-1}{5} \cdot \frac{9}{2}| = |-\frac{9}{10}| = \frac{9}{10}$$

The reader should note that the average quantity demanded of 20 units and the average price of \$4.50 per unit (= 9/2) is in Figure 5.3 at the midpoint between the two points on the demand curve that we have been considering. Another point to notice is that it is possible to conclude that demand is inelastic in this case because 9/10 is less than 1. Therefore, consumers are relatively unresponsive to price changes, at least between these two points on the demand curve.

The arc elasticity formula works well when we wish to calculate the price elasticity of demand between two points on the demand curve as in Figure 5.3. At other times, we might wish to calculate the price elasticity of demand at a single point on the demand curve. It might seem like an impossible task given that our initial formula depends upon changes in price and quantity demanded. It is still possible, however, to calculate the price elasticity of demand in such situations if we know that the demand curve is linear and we can determine the slope. For example, if we have the information given in Figure 5.4, then we can calculate the price elasticity of demand at a specific point.

Figure 5.4: A Situation that Requires the Point Elasticity Formula



To understand how this information can be used to calculate the price elasticity of demand, we can convert our original formula in the following way:

$$E_D = \frac{\%\Delta Q^D}{\%\Delta P} = \frac{\frac{\Delta Q}{Q}}{\frac{\Delta P}{P}} = \frac{\Delta Q}{Q} \cdot \frac{P}{\Delta P} = \frac{P}{Q} \cdot \frac{\Delta Q}{\Delta P} = \frac{P}{Q} \cdot \frac{1}{\frac{\Delta P}{\Delta Q}}$$

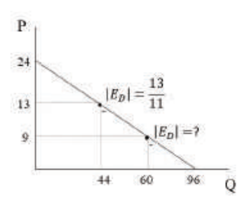
The final expression is the **point elasticity** formula. It is now possible to calculate the price elasticity of demand for a given P and Q on the demand curve. The reader should also note that ΔP/ΔQ represents the slope of a linear demand curve. If we use the information in Figure 5.4 to calculate the absolute value of the price elasticity of demand using the point elasticity formula, then we obtain the following result:

$$|E_D| = |\frac{P}{Q} \cdot \frac{1}{\frac{\Delta P}{\Delta Q}}| = |\frac{13}{44} \cdot \frac{1}{\frac{-24}{96}}| = |\frac{13}{44} \cdot \frac{1}{\frac{-1}{4}}| = |\frac{13}{44} \cdot -4| = \frac{13}{11}$$

In this case, the price elasticity of demand is greater than 1. Therefore, we conclude that demand is elastic at this point on the demand curve. That is, consumers are relatively responsive to the price change.

The reader might expect the price elasticity of demand to be the same at every point on this linear demand curve. After all, the slope is constant at every point on the demand curve. This conclusion is incorrect, however, as can be shown by calculating the price elasticity of demand at another point on the demand curve. For example, we might calculate the price elasticity of demand at a second point on the linear demand curve we just considered as shown in Figure 5.5.

Figure 5.5: The Decline in Price Elasticity of Demand along a Linear Demand Curve



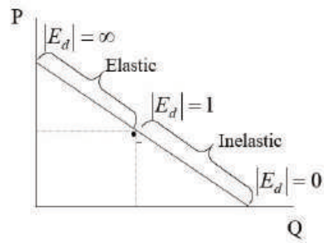
If we calculate the price elasticity of demand when the price is equal to $9 per unit and the quantity demanded is equal to 60 units, then we obtain the following result:
$|E_D| = |\frac{P}{Q} \cdot \frac{1}{\frac{\Delta P}{\Delta Q}}| = |\frac{9}{60} \cdot \frac{1}{\frac{-24}{96}}| = |\frac{9}{60} \cdot \frac{1}{\frac{-1}{4}}| = |\frac{9}{60} \cdot -4| = \frac{36}{60} = \frac{3}{5}$

In this case, we conclude that demand is inelastic because the price elasticity of demand is less than 1. It should also be noted that the slope of -1/4 remains the same in this calculation. The only change in this calculation compared with the calculation when the price is equal to $13 per unit is the specific P and Q that we use as we move down the demand curve. What we observe is that the price elasticity of demand declines as we move down the linear demand curve. This general result can be explained intuitively. That is, consumers are less responsive to price changes at lower prices than at higher prices. In other words, a 5% price reduction when the price is very high has a greater impact on the consumer's quantity demanded than a 5% price reduction when the price is very low.

We thus obtain a major result: *the price elasticity of demand falls as the price falls along a linear demand curve.*

The point elasticity formula can also be used to divide a linear demand curve into elastic and inelastic sections as shown in Figure 5.6.

Figure 5.6: The Division of a Linear Demand Curve into Elastic and Inelastic Sections



For example, we can calculate the price elasticity of demand where the demand curve intersects the vertical axis as follows:

$$|E_D| = |\frac{P}{0} \cdot \frac{1}{\frac{\Delta P}{\Delta Q}}| = |-\infty| = \infty$$

In this case, for any price and slope, the price elasticity of demand will be (or will approach) infinity.

We can also calculate the price elasticity of demand where the demand curve intersects the horizontal axis as follows:

$$|E_D| = |\frac{0}{Q} \cdot \frac{1}{\frac{\Delta P}{\Delta Q}}| = 0$$

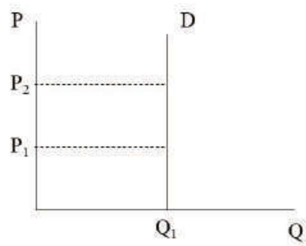
In this case, for any Q and slope, the price elasticity of demand will be zero.

Furthermore, the point elasticity formula indicates that as the price falls and the quantity demanded rises along a linear demand curve, the ratio of P to Q (that is, P/Q) will decline causing the price elasticity of demand to fall. Therefore, the price elasticity of demand falls continuously from its highest value of infinity to its lowest value of zero as we move down along the linear demand curve. Due to the continuous decline in the price elasticity of demand, at some point, the price elasticity of demand will equal 1. That is, demand will be unit elastic. This point marks the separation between the elastic portion of the demand curve (where $|E_D|>1$) and the inelastic portion of the demand curve (where $|E_D|<1$).

Cases of Extreme Elasticity

In certain situations, demand may become perfectly elastic or perfectly inelastic at every price on the demand curve. These situations are unusual but theoretically possible. For example, a consumer might become perfectly unresponsive to price increases if she is in desperate need of a life-saving medication. No matter how high the price rises, the consumer will pay the price and not reduce her quantity demanded. Of course, even this case has limits because the consumer will eventually become *unable* to make the purchase even if she is still *willing* to do so. (The reader will recall that both ability and willingness to pay are required for the demand for a product to exist.) Nevertheless, in this case, the demand curve becomes perfectly vertical as shown in Figure 5.7.

Figure 5.7: A Perfectly Inelastic Demand Curve



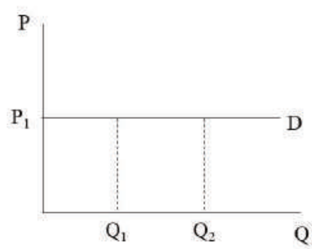
As the reader can see, even when the price rises from P1 to P2, the quantity demanded remains unchanged at Q1. Because the slope of the demand curve is infinite in this case, the use of the point elasticity formula yields the following result:

$$|E_D| = |\frac{P}{Q} \cdot \frac{1}{\frac{\Delta P}{\Delta Q}}| = |\frac{P}{Q} \cdot \frac{1}{\infty}| = 0$$

It is clear then that demand is perfectly inelastic when the slope of the demand curve is infinite. It should also be clear that the slope of the demand curve and the price elasticity of demand are inversely related.

A second extreme case arises when consumers become perfectly responsive to price changes. That is, the smallest rise in price may lead consumers to reduce their quantity demanded to 0. Alternatively, the smallest reduction in price might lead consumers to increase their quantity demanded to infinity (or a very large quantity, at least!). This might occur when a consumer knows that a perfectly good substitute product is available at the same price. As soon as the price of the product the consumer is buying rises at all, he switches immediately to the other product and his quantity demanded of the first product falls to zero. This situation is depicted in Figure 5.8 as a horizontal demand curve.

Figure 5.8: A Perfectly Elastic Demand Curve



As the reader can see, for any price above P_1 the quantity demanded will equal zero and for any price below P_1 the quantity demanded will soar to infinity. Because the slope of the demand curve is zero in this case, the price elasticity of demand can be calculated as follows:

$$|E_D| = |\frac{P}{Q} \cdot \frac{1}{\frac{\Delta P}{\Delta Q}}| = |\frac{P}{Q} \cdot \frac{1}{0}| = \infty$$

Demand is perfectly elastic then when the slope of the demand curve is equal to 0, and we again see that the slope of the demand curve and the price elasticity of demand are inversely related.

Key Determinants of the Price Elasticity of Demand

At this point, we have considered several ways to calculate the price elasticity of demand, but it is also necessary to consider the factors that cause the demand for a product to be more (or less) elastic. The key determinants of the price elasticity of demand include the following:

The nature of the good: Is it a luxury or a necessity?

If the good is a luxury item then consumers can do without it. If the price rises, consumers will be highly responsive to the price change. Similarly, if the price falls, consumers will be eager to enter this market. On the other hand, if the good is a necessity, like the life-saving medication described above, then consumers will not be able to reduce their purchases very much when the price rises. Alternatively, if the price falls, consumers will not expand their purchases much because they were already purchasing the amount required before the price fell. In general, the price elasticity of demand for luxuries tends to be greater than the price elasticity of demand for necessities, other things the same.

The budget share devoted to the purchase of the good

If consumers spend a large percentage or share of their budgets on a good, then the demand for the good tends to be more elastic, other things the same. The reason is obvious. Consumers will feel a greater pinch from a rise in the price of an automobile than they will feel from the rise in the price of chewing gum. Consumers of automobiles are, therefore, more responsive to price changes than consumers of chewing gum.

The time span during which the good is purchased

The reader should recall from Chapter 3 that the quantity demanded of a product is a flow variable. That is, it is defined per period, such as a day, a week, a month, or a year. If the period is very short, such as a day, then the demand for a good tends to be inelastic, other things the same. For example, when the price of gasoline rises sharply, consumers are often slow to reduce their purchases. Consumers still need to drive to work and to school. Over long periods, however, consumers can begin to seek out substitute forms of transportation. For example, they can carpool, ride the bus, or ride a bike. As a result, consumers are more responsive to price changes over longer periods than over shorter periods. In general, the demand for a good is more elastic in the long run than in the short run, other things the same.

The existence of substitutes

Finally, if many close substitutes for a good exist, then consumers tend to be more responsive to price changes. For example, if the price of apple juice rises sharply, consumers can substitute towards grape juice. As a result, the quantity demanded of apple juice drops sharply. On the other hand, if the price of juice in general rises, then consumers can substitute towards milk or soda, but these alternatives are not very good substitutes for juice. As a result, the demand for juice is much more inelastic than the demand for apple juice. In general, when more close substitutes for a good exist, demand tends to be more elastic, other things the same.

Below are a few examples of demand elasticities in different industries.[2]

Example 1: The price elasticity of demand for restaurant meals has been estimated to be 2.27.

Example 2: The short run price elasticity of demand for gasoline has been estimated to be 0.3.

Example 3: The price elasticity of demand for premium white pan bread has been estimated at 1.01.

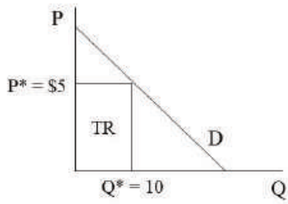
The high demand elasticity of restaurant meals can be attributed to the perception among many consumers that they are luxury goods. The low demand elasticity for gasoline can be attributed to the short period and the perception that gasoline is a necessity. Finally, the relatively elastic demand for premium white pan bread may be attributed to the existence of many close substitutes.

The Relationship between the Price Elasticity of Demand and Sales Revenue

When a firm sells a particular product, its sales revenue, or total monetary receipts, depends on two important variables: the price per unit and the number of units sold. In other words, a firm's revenue depends on price and quantity demanded. Specifically, we can calculate a firm's revenue by multiplying the product price by the quantity demanded. If we consider a linear demand curve, like the one shown in Figure 5.9, then it is possible to represent **total revenue** (TR) as the product of P and Q, or as the area of the box under the demand curve.

$$TR = PQ$$

Figure 5.9: The Demand Curve and Total Revenue



In this case, the total revenue is $50 or $5 per unit times 10 units.

A manager of a firm now faces a difficult task. If the manager aims to increase revenue, it might seem logical to raise the price of the good. The problem is that the quantity demanded will fall as the price rises, according to the law of demand. Therefore, we observe two conflicting effects on total revenue when the price changes, as shown below, leaving the overall effect an open question.

$$\overset{?}{TR} = \overset{\uparrow}{P}\overset{\downarrow}{Q}$$

The manager must ask which factor has the greater impact on revenue. In other words, what is the net effect of a change in price?

The answer to this question, it turns out, depends on the price elasticity of demand. Let's consider a price reduction that leads to a rise in quantity demanded. Although we do not provide a rigorous proof of this result, we can provide an intuitive explanation for these changes. If the price falls by the same percentage as the rise in quantity demanded, then the net effect on total revenue is zero. On the other hand, if the price falls by a greater percentage than the rise in quantity demanded, then the net effect on total revenue is negative.[3] Finally, if the price falls by a smaller percentage than the rise in quantity demanded, then the net effect on total revenue is positive. These results and the implications for demand elasticity are shown below with larger arrows indicating larger percentage changes.[4]

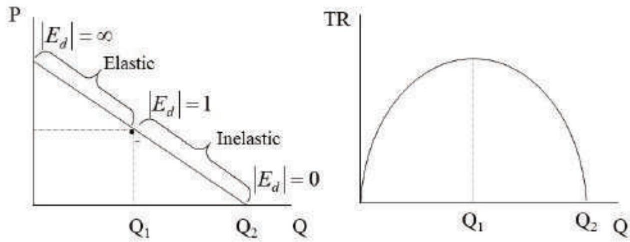
$$P_{\downarrow}\ and\ Q \uparrow\ then\ TR \uparrow \Rightarrow Demand\ is\ elastic$$

$$P \downarrow\ and\ Q \uparrow\ then\ \overline{TR} \Rightarrow Demand\ is\ unit\ elastic$$ image

$$P \downarrow\ and\ Q_{\uparrow}\ then\ TR \downarrow \Rightarrow Demand\ is\ inelastic$$ image

Because we identified the elastic and inelastic portions of a linear demand curve earlier in this chapter, we can now use that information to determine the shape of the total revenue curve. Figure 5.10 reveals that as the price falls from its peak, total revenue rises because demand is elastic. Once we reach the point where demand is unit elastic, total revenue does not change (the meaning of the overbar above TR in the unit elastic case shown above). Finally, if price falls far enough, then total revenue declines because demand is inelastic.

Figure 5.10: The Relationship Between Elasticity and Revenue



The reader can see that when the price elasticity of demand is infinite, total revenue is equal to 0. Although the price is high, the quantity demanded is zero. Similarly, when the price is set at 0, total revenue is equal to 0 even though the quantity demanded is high at Q_2. In between these two extremes, the total revenue rises and then falls as the price declines. Because total revenue rises prior to the point of unit elasticity and falls after the point of unit elasticity, it follows that total revenue reaches its peak at the point of unit elasticity. Although total revenue reaches its maximum at Q_1, the reader should not assume that this level of output is the optimal choice for the firm. The firm must also consider production cost, which is a concept that is discussed at great length in Chapter 7.

Other Measures of Elasticity: Supply, Cross-Price, and Income

The price elasticity of demand is a widely used concept among neoclassical economists, in part because of its connection to sales revenue. Several other elasticity concepts are also useful, namely the price elasticity of supply, the cross-price elasticity of demand, and the income elasticity of demand. This section only provides a brief overview of these concepts.

The Price Elasticity of Supply

Just as consumers are responsive to changes in price, so are sellers. To measure the responsiveness of sellers to price changes, neoclassical economists use the **price elasticity of supply** defined below:

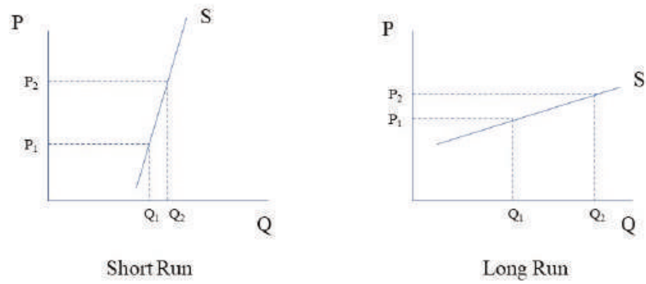
$$E_S = \frac{\%\Delta Q^S}{\%\Delta P}$$

The formula for the price elasticity of supply (E_S) measures the percentage change in the quantity supplied for a given percentage change in price. For example, if the price of a product increases by 8% and the quantity supplied rises by 24%, then the price elasticity of supply equals 3. It should be noted that it is not necessary to calculate the absolute value of the price elasticity of supply as we did earlier with the price elasticity of demand. The reason is that the price elasticity of supply will virtually always be a positive number due to the law of supply. That is, price increases will lead to increases in quantity supplied and price reductions will lead to reductions in quantity supplied.

When the price elasticity of supply equals 3, it may be interpreted to mean that a 1% rise in price leads to a 3% rise in quantity supplied, or a 1% reduction in price leads to a 3% reduction in quantity supplied. The interpretations of price elasticity of supply are also like the interpretations of the price elasticity of demand. For example, if the price elasticity of supply equals 1, then supply is unit elastic. If it is greater than 1, then supply is elastic. If it is less than 1, then supply is inelastic. If it is infinite, then supply is perfectly elastic. If it is equal to 0, then supply is perfectly inelastic.

The primary determinant of the price elasticity of supply is the time frame. If the price of apples rises in a very short period such as a month, for example, then producers of apples cannot expand production very quickly. They can hire more workers to pick more apples, but they cannot grow more apples in a month's time. The price elasticity of supply in the short run tends to be very low, as a result. Over a longer period, it may be possible to purchase additional orchards or even to grow more trees. Supply then becomes much more elastic. That is, producers can expand production as the price rises over long time periods. Figure 5.11 shows an inelastic short run supply curve and a much more elastic long run supply curve.

Figure 5.11: Price Elasticity of Supply in the Short Run and Long Run



The Cross-Price Elasticity of Demand

The **cross-price elasticity of demand** (or simply, cross elasticity) allows us to measure the percentage change in the quantity demanded of a product for a given percentage change in the price of a different product. For example, the cross elasticity of demand between goods A and B can be written as follows:

$$E_{A,B} = \frac{\%\Delta Q_A^D}{\%\Delta P_B}$$

If the price of good B rises by 4% and the quantity demanded of good A rises by 2%, then the cross elasticity is equal to +1/2. In this case, the sign is important. The fact that the consumer purchases more of good A when the price of good B rises indicates that the consumer considers good A to be a **substitute** for good B. Similarly, if the price of peanut butter rises by 6% and the quantity demanded of jelly falls by 2%, then the cross elasticity is

-1/3. Because the consumer reduces purchases of jelly when the price peanut butter rises, it suggests that the goods are often consumed together. In other words, the two goods are **complements**. Finally, the two goods are regarded as **unrelated goods** if the quantity demanded does not change at all when the price of the other good changes. That is:

$$E_{A,B} = \frac{\%\Delta Q_A^D}{\%\Delta P_B} > 0 \Rightarrow \ Substitutes$$

$$E_{A,B} = \frac{\%\Delta Q_A^D}{\%\Delta P_B} < 0 \Rightarrow \ Complements$$

$$E_{A,B} = \frac{\%\Delta Q_A^D}{\%\Delta P_B} = 0 \Rightarrow \ Unrelated\ Goods$$

The reader should notice that a positive value for the cross elasticity may be obtained either with a +/+ or a -/-. Similarly, a negative value for the cross elasticity may be obtained either with a +/- or a -/+.

The cross elasticity has an important application in antitrust law. Consider two firms that are trying to initiate a merger. If the merger reduces competition and leads to higher prices, then the merger will harm consumers. Antitrust laws are in place to prevent such mergers, but it is not always obvious whether two goods are substitutes for one another. If they are, then the merger will reduce competition. As a result, the U.S. Justice Department, which enforces the nation's antitrust laws, may ask an expert to testify in antitrust cases that two goods have a positive cross elasticity. Such evidence supports the claim that the goods are substitutes and that a merger of the two firms will harm American consumers.

The Income Elasticity of Demand

One final elasticity measure that we will consider is the

income elasticity of demand. This elasticity concept measures the responsiveness of consumers to a change in income. That is, it measures the percentage change in the quantity demanded of a product given a percentage change in income (Y). It can be written as follows:

$$E_Y = \frac{\%\Delta Q^D}{\%\Delta Y}$$

For example, if a 3% rise in income leads to a 12% rise in quantity demanded, then the income elasticity of demand is equal to +4. The sign of the income elasticity of demand is also important. The positive sign in this case indicates a positive relationship between income and quantity demanded. In Chapter 3, we explained that such goods are called **normal goods**. Alternatively, if a 3% rise in income leads to a 6% drop in quantity demanded, then the income elasticity of demand is -2. Such goods are called **inferior goods** because a negative relationship exists between income and quantity demanded. Finally, if the quantity demanded does not change at all when income changes, then the income elasticity of demand is equal to 0. Such goods are said to be **neutral** with respect to income. That is:

$$E_Y = \frac{\%\Delta Q^D}{\%\Delta Y} > 0 \Rightarrow \ Normal\ Goods$$

$$E_Y = \frac{\%\Delta Q^D}{\%\Delta Y} < 0 \Rightarrow \ Inferior\ Goods$$

$$E_Y = \frac{\%\Delta Q^D}{\%\Delta Y} = 0 \Rightarrow \ Neutral\ Goods$$

The reader should note that a positive income elasticity of demand may be obtained with either a +/+ or a -/-. The reader should also note that a negative income elasticity of demand may be obtained with either a +/- or a -/+.

Do Marxian Economists Use Elasticity Measures?

Before concluding this chapter, it is worth noting that Marxian economists do not make much use of elasticity measures. One reason is that the price elasticities of supply and demand are measures that relate to market supply and demand curves. For neoclassical economists, supply and demand provide the best explanation of market prices. For Marxian economists, supply and demand influence market prices, but these laws are subordinate to the law of value. That is, socially necessary abstract labor time governs commodity values and so it is the law of value that receives the attention of Marxian economists. Similarly, Marxian economists are interested in the class dynamics of capitalist societies. How consumers respond, and to what extent they respond, to price changes is not a major concern of Marxian economists. A second reason is that the two primary measures that Marxian economists use to draw comparisons across industries are unit-free measures. Recall that the rate of surplus value and the rate of profit measure the degree of exploitation of labor-power and the rate of self-expansion of capital in different industries, respectively. Both measures are already expressed in percentage terms and so direct comparisons across industries are possible. In Chapter 8, we will see that such comparisons across industries play an important role in the Marxian theory of industrial competition.

Following the Economic News

According to a 2011 study of Australian meat consumption, the domestic demand for meat is relatively elastic. Researchers found that all types of meat, as well as rice, margarine, and preserved vegetables, had elastic demands. On the other hand, the demands for milk, bread, fresh fruit, and fresh vegetables were relatively inelastic. In addition, Australian meat demand was found to be more elastic than meat demand in other nations, suggesting that "Aussies are more likely to reduce the amount of meat purchased or substitute one meat type for another if the price of their preferred choice goes up."[5]

Summary of Key Points

1. The price elasticity of demand measures the responsiveness of consumers to price changes.
2. Unlike the slope of the demand curve, the price elasticity of demand is a unit-free measure.
3. A larger absolute value of the price elasticity of demand implies that demand is more elastic.
4. The arc elasticity formula is used to calculate price elasticity of demand when moving between two points on a demand curve; the point elasticity formula is used to calculate the price elasticity of demand at a single point on the demand curve.
5. The price elasticity of demand depends on the nature of the good, the time span, the budget share, and the existence of substitutes.
6. Whether total revenue rises, falls, or remains the same when price changes, depends on whether demand is elastic, inelastic, or unit elastic.
7. The price elasticity of supply measures the responsiveness of sellers to price changes.
8. The cross-price elasticity of demand measures the responsiveness of consumers to a change in the price of a different good.
9. The income elasticity of demand measures the

responsiveness of consumers to a change in their incomes.

10. We do not use the absolute value of supply elasticity, cross elasticity, and income elasticity because the signs of these values are highly significant.

List of Key Terms

Price elasticity of demand

Elastic

Unit elastic

Inelastic

Perfectly inelastic

Perfectly elastic

Arc elasticity

Point elasticity

Total revenue

Price elasticity of supply

Cross-price elasticity of demand

Substitutes

Complements

Unrelated goods

Income elasticity of demand

Normal goods

Inferior goods

Neutral goods

Problems for Review

1. Suppose the price of a good rises by 6% and the quantity demanded falls by 30%. What is the price elasticity of demand (in absolute value terms)? Also, interpret your answer.
2. Suppose the price of a good falls from $62 to $58 and the quantity demanded rises from 1,760 units to 1,910 units. Calculate the price elasticity of demand using the arc elasticity formula. Also, interpret your answer.
3. Suppose the price of a good is $45 and the quantity demanded is 500 units. If the slope of the linear demand curve is -3, then what is the price elasticity of demand, according to the point elasticity formula? Also, interpret your answer.
4. If the price of gasoline rises and the total revenue received by sellers decreases, then what can you conclude about the relationship between the percentage change in quantity demanded relative to the percentage change in price (in absolute value)? What can you conclude about the price elasticity of demand?
5. If the price of cereal rises by 2% and the quantity demanded of milk falls by 0.5%, then what is the cross elasticity between the two products? Also, interpret your answer. Is your interpretation consistent with your expectations?
6. If your income rises by 10% and your quantity

demanded of used goods falls by 20%, then what is the income elasticity of demand? Interpret your answer. Is your answer consistent with your expectations?

Notes

1. I wish to acknowledge Prof. John Hoag for first emphasizing to me the units in which slope is measured.
2. Keat et al. (2013), p. 84-85.
3. Recalling that the $\%\Delta xy \approx \%\Delta x + \%\Delta y$, we can write that $\%\Delta PQ \approx \%\Delta P + \%\Delta Q$. If revenue falls, then $\%\Delta P + \%\Delta Q < 0$, and $\%\Delta Q < -\%\Delta P$. In the case of a price decrease and an increase in quantity demanded, it follows that the percentage change in price exceeds the percentage change in quantity demanded (in absolute value terms). Hence, demand is inelastic.
4. It is left to the reader to consider how to analyze the three cases when the price increases.
5. "Study Reveals Elastic Domestic Demand for Meat." *Beef Central*. October 25, 2011.

Chapter 6

THEORIES OF UTILITY MAXIMIZATION

Goals and Objectives:

In this chapter, we will do the following:

1. *Outline* the history of utility theory
2. *Describe* the traditional and modern neoclassical theories of utility maximization
3. *Explain* how to derive the demand curve using utility maximization
4. *Explore* criticisms of the neoclassical measure of welfare
5. *Learn* two competing solutions to the "paradox of value"
6. *Analyze* the complications that arise when preferences are endogenous

The Utilitarian Roots of the Theories

The theories that we will explore in this chapter have their roots in a philosophy called **utilitarianism** that developed in the nineteenth century. Utilitarianism is primarily a moral philosophy. According to its British founder, Jeremy Bentham (1748-1832), we ought to judge all social and economic policies according to the consequences that they carry for human happiness. If policies promote human happiness then they are morally right. If they diminish human happiness, then they are morally bankrupt. In other words, we ought to pursue the greatest happiness of the greatest number. Utilitarianism is mostly a *normative* theory, as defined in chapter 1. Its focus is primarily on what ought to be. John Stuart Mill is another famous thinker whose ideas on economics and politics utilitarianism greatly influenced, and the theory continues to have great influence among moral philosophers as well as economists.

By the late nineteenth century, many economists were beginning to reject the labor theory of value, arguably because its implications served critics of market capitalism more than it served its defenders. The notion that workers might be creating more value in production than they received in compensation was too threatening to those who wished to maintain the present social order. In its place, they discovered a theory of value that appeared to shift the focus back to the value that consumers obtain from consumption and away from labor content or cost of production. It was this discovery that gave rise to a school of economic thought that was so far removed from the classical economics of Smith and Ricardo that it was given the new title of neoclassical economics.

In the 1870s, three economists working independently co-discovered a concept that would form the basis of neoclassical microeconomics for the next century and a half. Leon Walras in France, William Stanley Jevons in Britain, and Carl Menger in Austria each developed the concept of **marginal utility** and identified an economic law to which they claimed all consumers are subject. They shifted the emphasis from the *normative* idea inherent in traditional utilitarianism (that we should pursue the greatest happiness of the greatest number) to a *positive* idea that economic agents, acting in a world of scarce resources, seek to maximize their own happiness. As economic agents go about maximizing their happiness, the economic law to which each is subject governs and determines their behavior in the marketplace.

The Concept of Utility in the Traditional Theory of Utility Maximization

Before we define this economic law with precision, we must first define the concept of **utility** more carefully. The early neoclassical economists aimed to shift the focus of value theory back to the consumption side, but they did not wish to use the classical notion of use value. The use value of a commodity refers to the useful properties of a good or service. Utility, on the other hand, refers to the *satisfaction* that the user experiences in consumption.

One reason neoclassical economists define utility in terms of satisfaction is that they recognize that the experience of consumption is highly subjective. How useful the properties of a good or service are perceived to be depends on the user's subjective mental experience. For that matter, an item might seem to an external observer to be useless. For example, in 1975 an entrepreneur had the idea to sell ordinary rocks as pets. These Pet Rocks had no function at all. They were simply good for a laugh. Nevertheless, the owner of the business became a millionaire before the item fell out of fashion. In addition, the good or service might not always appear to lead to pleasure. For example, if you go to the doctor for a shot, it might be painful rather than pleasurable. Nevertheless, on some level, you recognize that it is good for you to have it, and so you receive satisfaction despite the sting.

One of the strangest properties of utility according to the traditional theory of utility maximization is its quantifiable nature. Early neoclassical economists assumed that they could measure utility in terms of a standardized unit called **utils**. For example, you might obtain 25 utils of satisfaction from eating a bowl of cereal. This claim might seem silly, but remember that it is only an assumption for model building. Early neoclassical economists did not literally believe that satisfaction was a peculiar substance in the brain. Instead, they regarded it as a useful theoretical tool that would yield useful results. In other words, if we assume that satisfaction is measurable, does that assumption help us to *explain* and *predict* the behavior of individual consumers? Early neoclassical economists believed that it was so.[1]

The Law of Diminishing Marginal Utility

We next introduce a very important distinction between two key variables in the traditional theory of utility maximization: **total utility** and **marginal utility**. Total utility (TU) refers to the overall amount of satisfaction derived from the consumption of a good or service. For a certain period, we measure TU in utils. Marginal utility (MU) refers to the *additional* satisfaction obtained from

the consumption of an *additional* unit of a good or service. We measure MU in utils per unit as follows:
$MU = \frac{\Delta TU}{\Delta Q}$

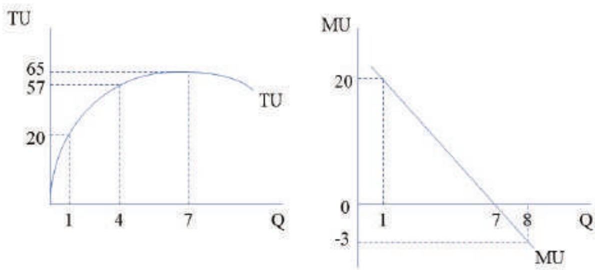
For example, six mozzarella sticks might give you a total utility of 65 utils, but the third mozzarella stick might only give you a marginal utility of 12 utils for that unit consumed. More generally, as an individual consumes more of a good, a clear pattern begins to emerge, argues the neoclassical economist. Table 6.1 shows what might happen to the total utility and marginal utility of a consumer as she consumes more mozzarella sticks.

Table 6.1: Total Utility and Marginal Utility

Mozzarella Sticks	Total Utility (TU)	Marginal Utility (MU)
0	0	--
1	20	20
2	35	15
3	47	12
4	57	10
5	62	5
6	65	3
7	65	0
8	62	-3

As the table indicates, as the consumer consumes more mozzarella sticks, the TU rises quite a bit even though it eventually begins to decrease. MU, on the other hand, falls from the beginning and eventually becomes negative. We can also graph these relationships as in Figure 6.1 where only a few of the points on each curve have been included.

Figure 6.1: Total Utility and Marginal Utility



These graphs of the TU and MU curves offer two different ways to visualize the **law of diminishing marginal utility** to which all individual consumers are subject, according to neoclassical economists. The law of diminishing marginal utility states that marginal utility declines as a consumer consumes more of a good. This reduction in MU is the result of the consumer attempting to satiate herself due to her consumption of the good. The marginal utility of each additional unit consumed is smaller and smaller until eventually an additional unit makes a negative addition to the total utility.

The second way of stating the law of diminishing marginal utility is in terms of total utility. Figure 6.1 shows that total utility rises as more of the good is consumed but at a decreasing rate. Eventually, total utility begins to fall as the consumer approaches satiation due to her consumption of the good.

It is worth noting the relationship between the two graphs. Mathematically, the slope of the TU curve is MU ($=\Delta TU/\Delta Q$). It should come as no surprise then that when total utility reaches its maximum, marginal utility equals zero utils per unit. That is, the slope of the TU curve is zero at the maximum point on the TU curve.

The law of diminishing marginal utility is perfectly compatible with the law of demand. If a consumer experiences a reduction in marginal utility as he consumes more of a good, then he will need a lower price to justify the purchase of an additional unit. That is, he will only be willing to purchase another unit if the price falls because that additional unit offers less satisfaction than the previous unit he consumed. Interestingly, the law of diminishing marginal utility also appears to be compatible with the discovery we made in chapter 5 that the price elasticity of demand falls all the way down a linear demand curve. It makes sense that a consumer would be less responsive to price changes at lower prices because those units correspond to lower marginal utilities.

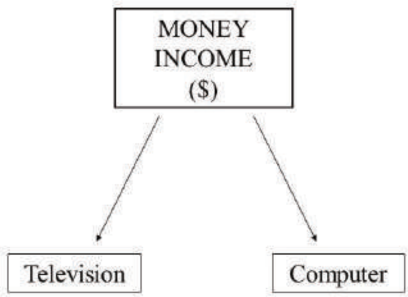
The Traditional Theory of Utility Maximization

The law of diminishing marginal utility tells us how each consumer responds to the increased consumption of a good, but it does not tell us much about the consumer's choices. Consumers generally consume more than one good. Consumers also must purchase these goods at constant prices and with limited incomes. To understand how much of each good a consumer will choose to purchase, we need a model to assist us with the analysis. To create an environment in which a single consumer makes decisions about how much of each good to purchase, we will assume that the consumer aims to maximize her total utility. It is also assumed that the consumer has complete preferences for the two goods available in the market and is subject to the law of diminishing marginal utility. Finally, the consumer faces a constant money income and constant prices for the two goods.

The statement that a consumer has complete preferences means that he knows exactly how many utils he will obtain from the consumption of a specific quantity of a good or service. In addition, the reader should notice the neoclassical entry point of given individual preferences and a given resource (or income) endowment. Overall, these assumptions are enough to make possible an analysis of the consumer's utility maximizing choice.

We should begin with a simple example involving two goods. The consumer is trying to decide between the purchase of a television and a computer. Figure 6.2 captures the economic problem facing this consumer.

Figure 6.2: The Economic Problem Facing the Consumer



If the television will give the consumer 400 utils and the computer will give the consumer 800 utils, then one might conclude that the consumer should purchase the computer to maximize total utility. The problem with this line of reasoning is that it completely fails to consider the prices of the products. The price of the television might be $100 and the price of the computer might be $400. It turns out that the consumer will choose to purchase the television even though it gives the consumer less utility because the price is so much lower than the price of the computer.

To prove this point, we need to compare the marginal utility per dollar spent on each good (MU/$) by dividing the marginal utility of each good by its price. For example, the MU per dollar spent on the television is 400 utils/$100 or 4 utils per dollar spent. Similarly, the MU per dollar spent on the computer is 800 utils/$400 or 2 utils per dollar spent. Because the MU per dollar spent on the television is greater, the consumer will get more bang for her buck when she buys the television. We may state the general rule as follows:

The consumer will always purchase first the good that gives the higher MU per dollar spent.

We are now able to derive a more general result about utility maximizing behavior. Let us consider an example in which an individual is trying to decide how much of good X and how much of good Y to purchase. We will assume that the consumer's income (M) is \$35, the price of good X (P_X) is \$5 per unit, and the price of good Y (P_Y) is \$3 per unit. We also assume that the consumer has clear preferences and so he or she knows the TU and MU for each quantity of the two goods consumed. Table 6.2 represents this information.

Table 6.2: Utility Maximization in the Traditional Theory

Unit	TU_X	MU_X	MU_X/P_X	TU_Y	MU_Y	MU_Y/P_Y
0	0	--	--	0	--	--
1	25	25	5	45	45	15
2	45	20	4	69	24	8
3	60	15	3	84	15	5
4	70	10	2	93	9	3
5	75	5	1	99	6	2
6	78	3	0.60	96	-3	-1
7	79	1	0.20	90	-6	-2

As the table shows, the consumer knows her preferences, and the law of diminishing marginal utility holds for each of the goods. In both cases, the MU falls as the consumer consumes more of the good. To calculate MU per dollar spent, we simply divide each MU by the price of the good in question.

The next question we wish to ask is how much of each good the consumer will choose given the income and the prices of the two goods. Because the consumer aims to maximize utility, the consumer will always select first the good that gives her the highest marginal utility per dollar spent. For example, assume that the consumer has an empty shopping cart and is trying to decide whether to purchase the first unit of good X or the first unit of good Y. Because the MU per dollar spent on the first unit of Y (= 15) is so much larger than the MU per dollar spent on X (= 5), the consumer will buy the first unit of Y. Next, the consumer must decide between the second unit of Y and the first unit of X. Because the eight utils per dollar spent on the second unit of good Y exceed the five utils per dollar spent on good X, the consumer will purchase the second unit of good Y. We need to keep track of expenditure as we proceed. Thus far, the consumer has spent $6.

Next, the consumer must decide between the third unit of good Y and the first unit of good X. In this case, either choice leads to five utils per dollar spent. The consumer will be completely indifferent in this case between the third unit of Y and the first unit of X. The reader might think that the consumer should choose the third unit of Y because it is cheaper but that would be a mistake. We already considered the fact that Y is cheaper than X when we divided the MU of each unit of each good by the appropriate price. Let us then assume that the consumer will choose the first unit of good X, bringing total expenditure to $11. Now the consumer must choose between the second unit of good X and the third unit of good Y. Because 5 > 4, the consumer will choose the third unit of good Y, bringing total expenditure to $14.

The next choice is between the fourth unit of Y and the second unit of X. Since 4 > 3, the consumer chooses the second unit of X. Expenditure has now reached $19. Next, the consumer is indifferent between the third unit of X and the fourth unit of Y because each yields 3 utils per dollar spent. If the consumer buys the fourth unit of Y, then total expenditure is up to $22. Now the consumer compares the MU per dollar spent on the fifth unit of good Y with the MU per dollar spent on the third unit of good X. Since 3 > 2, the consumer chooses to buy X, bringing total expenditure up to $27. Finally, the consumer is indifferent between the fourth unit of X and the fifth unit of Y. If the consumer chooses the fifth unit of Y, then total expenditure is $30. With the remaining $5, the consumer can buy the fourth unit of X because the sixth unit of Y has a negative MU and would reduce total utility.

With that final purchase, the consumer has purchased five units of good Y and four units of good X. It is the utility-maximizing combination of the two goods. In addition, the consumer has spent the entire $35 of income. Because we have been assuming that only the consumption of goods and services generates utility, it follows that the consumer must spend all income. If some income is not spent (i.e., if saving occurs), then utility cannot be at its maximum. The reader might object that individuals do choose to save. Neoclassical economists do analyze saving from a microeconomic perspective, but it involves a utility maximizing choice between present and future consumption rather than between two present goods.[2]

One other major result that deserves special attention is the fact that the MU per dollar spent on goods X and Y are the same when the consumer is maximizing utility. The final unit of each good yields 2 utils per dollar spent. It is possible to construct an example where the marginal utilities per dollar spent on each good are only approximately equal due to the consumer running out of money before the equality is reached. Nevertheless, the point remains that the consumer will continue to move back and forth between the two goods until the marginal utilities per dollar spent are nearly, if not exactly, equal. Excluding possible exceptions, we can sum up as follows:

The consumer maximizes utility when the marginal utilities per dollar spent are the same for both goods and all income is spent.

Algebraically, it is possible to write this condition as follows:

$$\frac{MU_X}{P_X} = \frac{MU_Y}{P_Y} \; (all\ income\ is\ spent)$$

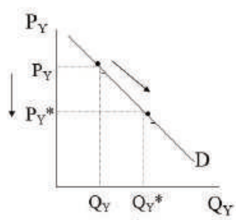
Furthermore, if all income is spent and $\frac{MU_X}{P_X} > \frac{MU_Y}{P_Y}$, then the consumer should buy more X and less Y. As more X is purchased, the MU of X falls due to diminishing marginal utility. Similarly, as less Y is purchased, the MU of Y rises due to diminishing marginal utility. These changes will continue until the marginal utilities per dollar spent are the same for the two goods.

Similarly, if all income is spent and $\frac{MU_Y}{P_Y} > \frac{MU_X}{P_X}$, then the consumer should buy more Y and less X. As more Y is purchased, the MU of Y falls due to diminishing marginal utility. Similarly, as less X is purchased, the MU of X rises due to diminishing marginal utility. These changes will continue until the marginal utilities per dollar spent are the same for the two goods.

The Traditional Derivation of the Individual Demand Curve

The utility maximizing condition described in the last section can be used to derive the individual demand curve. That is, we can gain a deeper understanding as to why the individual demand curve slopes downward when we think in terms of utility maximizing behavior. Suppose, for example, that the price of good Y declines. The law of demand states that the quantity demanded of Y will rise, other things the same. This situation is depicted in Figure 6.3.

Figure 6.3: The Derivation of the Individual Demand Curve



Assume that the consumer is maximizing utility at the original price of P_Y. In other words, the initial condition when all income is spent is the following:

$$\frac{MU_X}{P_X} = \frac{MU_Y}{P_Y}$$

Now suppose that P_Y declines to P_Y*. It follows that:

$$\frac{MU_X}{P_X} < \frac{MU_Y}{P_{Y*}}$$

Because the MU per dollar spent on good Y is greater than the MU per dollar spent on good X, the consumer will purchase more Y. Because the price of good Y fell and the quantity demanded of good Y rose, we can conclude that the demand curve slopes downward. It is worth noting that the increase in the quantity demanded of Y will eventually cease (when Q reaches Q_Y^*) because the diminishing marginal utility of Y will restore the equality between the two marginal utilities per dollar spent.

The Paradox of Value

The traditional theory of utility maximization also helps to resolve a **paradox** in the history of economic thought. A paradox is a statement that appears to be false but that contains an element of truth. The paradox of value refers to the fact that many goods, like diamonds, appear to be not very useful even though they have high prices. Similarly, the most useful goods, like water, often have the lowest prices.

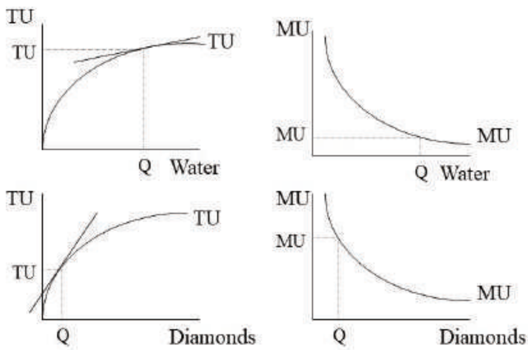
The classical economists, like Smith and Ricardo, believed that they had successfully found the solution in the labor theory of value. That is, diamonds have a high price because they require a lot of labor time for their production. Similarly, water has a low price because it requires very little labor time for its production.[3]

Early neoclassical economists in the 1870s offered a different solution to the paradox than that which the classical economists offered. Their discovery of the marginal utility concept allowed them to develop a different solution to the paradox that emphasized consumer satisfaction rather than production cost. Specifically, they pointed out that the proper resolution to the paradox requires the distinction between total utility and marginal utility. That is, goods that most people consider to be useful, like water, generate a great deal of total utility for consumers. The marginal utility of the last unit consumed, however, is very low due to the law of diminishing marginal utility and the fact that so many units are consumed. It is the marginal utility of the last unit consumed that governs its price, according to this solution. In other words, the low MU of water explains its low price.

On the other hand, goods that most people regard as relatively useless, like diamonds, do not give consumers a great deal of total utility. The marginal utility of the last unit consumed, however, is very high due to the law of diminishing marginal utility and the fact that relatively few units are consumed. In this case, it is the high MU of diamonds that explains its high price.

We can also graphically represent the marginal utility solution to the paradox of value as shown in Figure 6.4.

Figure 6.4: The Marginal Utility Solution to the Paradox of Value



The top graphs in the figure show that the total utility obtained from the consumption of water is very high due to the large volume consumed. The marginal utility of the last unit consumed is very low for that same reason. The bottom graphs in the figure show that the total utility obtained from the consumption of diamonds is very low because of the small amount consumed. For the same reason, the marginal utility of the last diamond consumed is very high. The reader should also observe the straight lines that are just tangent to the TU curves at the quantities consumed. The slopes of the tangent lines represent the marginal utilities. The relatively flat tangent line in the top left graph indicates a low marginal utility. The relatively steep tangent line in the bottom left graph indicates a high marginal utility.

Table 6.3 summarizes all the information in this section, including the paradox of value itself as well as the classical and early neoclassical solutions to the paradox.[4]

Table 6.3: The Paradox of Value

		Water	Diamonds
The Paradox	Price	Low	High
	Total Utility	High	Low
The Solutions	Embodied Labor	Low	High
	Marginal Utility	Low	High

Criticisms of the Neoclassical Measure of Welfare

The concepts of consumers' surplus and producers' surplus that are discussed in Chapter 3 are presented as being logically consistent with the rest of neoclassical microeconomic theory. The concepts, however, are subject to two serious criticisms that can only be understood once we have a firm grasp of the theory of utility maximization.

The first major criticism is that the concepts implicitly assume a constant marginal utility of money. For example, suppose that I am willing and able to pay $5 for an ice cream sundae, but I only need to pay the market price of $2. I enjoy a consumer surplus of $3. Suppose further that I purchase a second ice cream sundae for $2 and that I am willing and able to pay $4 for the second one. I enjoy a consumer surplus of $2 on the second one that I purchase. My total consumer surplus is $5 (=

$3+$2). This measure of my welfare only makes sense if the marginal utility of money is constant for me. That is, if the law of diminishing marginal utility applies to money, then the first dollar of consumer surplus that I receive will contribute more to my welfare than the last dollar that I receive. In other words, the yardstick by which we measure welfare must remain the same for consumers' surplus to be regarded as an internally consistent measure of consumers' welfare. The same criticism applies to producers' surplus.

The second major criticism is that the concepts also implicitly assume that the marginal utility of money is the same for all buyers and sellers. This criticism becomes relevant when we attempt to aggregate the surpluses of different individuals. For example, suppose that my consumer surplus is a total of $75 and your consumer surplus is also $75. If we add our surpluses together, we obtain a total of $150. What if my marginal utility of money is much higher than your marginal utility of money? In that case, my surplus should count more in the aggregate measure than your surplus, but in this case, our surpluses contributed equally to this measure of consumer welfare. Both criticisms raise serious problems that have been recognized by neoclassical economists. Despite the difficulties, the concepts are still used and students are rarely encouraged to critically reflect on the problems inherent in the use of this welfare measure.

The Modern Theory of Utility Maximization

As we have seen, the early neoclassical economists assumed that utility was quantifiable and measurable. When utility can be quantified and measured, it is called

cardinal utility. During the twentieth century, neoclassical economists realized their theory would be stronger if they abandoned an assumption as strict as this one. Their effort to rid consumer theory of this strict assumption is a good example of the application of a philosophical principle known as Ockam's Razor. According to this principle, the strongest theory is the one based on the simplest assumptions, other things the same.

This principle has been applied in many contexts, including in the natural sciences. For example, during the nineteenth century, Charles Darwin developed the theory of natural selection to explain how organisms adapt to their natural environments over long periods of time. According to Darwin, random genetic mutations cause variations in the physical characteristics among the members of a species. Those members that develop the most advantageous characteristics for survival will thrive and reproduce. Those members with the least advantageous characteristics will die and fail to reproduce. As a result, the most advantageous characteristics are passed down and the least advantageous characteristics disappear from the species.

A little-known fact among non-scientists is that a competing theory of evolution also existed in the nineteenth century. According to a French anatomist by the name of Lamarck, the members of a species acquire advantageous characteristics during their lifetimes in response to their specific environments. For example, a giraffe stretches its neck to reach leaves further up on a tree, and thus acquires a longer neck, which it then passes down to its offspring. This claim, that an organism acquires traits that alter the genetic material passed down to its offspring, is an assumption of Lamarck's theory that is much more complicated than Darwin's assumption that random genetic mutations are the source of the advantageous characteristics. Because Darwin's theory was based on a simpler set of assumptions, it was adopted and Lamarck's theory was rejected. It should be emphasized that Darwin's theory was not *proven* once and for all. It was accepted as the best theory due to the application of Ockam's Razor.

Similarly, neoclassical economists aimed to replace the less plausible assumption of quantifiable satisfaction with a simpler assumption. That is, they rejected cardinal utility and replaced it with a different kind of utility called **ordinal utility** in their models of consumer behavior. The ordinal utility concept implies that consumers receive satisfaction from the consumption of goods, but they are only able to create a ranking among alternative commodity bundles (i.e., combinations of the goods).

For example, bundle 1 may consist of 2 apples and 1 banana whereas bundle 2 may consist of 1 apple and 2 bananas. If we use the notion of ordinal utility, then a consumer can only say one of three things: 1) Bundle 1 is preferred to bundle 2; 2) bundle 2 is preferred to bundle 1; or 3) bundles 1 and 2 are equally preferred (i.e., the consumer is indifferent between the two bundles).

On the other hand, if we use the concept of cardinal utility, then it would be possible to state that bundle 1 creates 10 utils for the consumer and bundle 2 creates 5 utils for the consumer. In that case, bundle 1 is not only preferred to bundle 2, but it is exactly twice as preferred as bundle 2. That is, the cardinal utility assumption allows us to state *how much* more preferred one bundle is than another. With the ordinal utility assumption, it is only possible to state that one bundle is more preferred than another, but it is impossible to state how much more preferred that bundle is.

The Implicit Ideological Significance of the Ordinal Utility Assumption

Before we explore the modern theory of utility maximization at greater length, we should pause to consider the implicit ideological significance of the ordinal utility assumption. It will not be immediately obvious to the uninitiated reader what role this assumption plays in neoclassical theory, but it can be explained easily enough if one is willing to be direct about it.

If cardinal utility is assumed, then satisfaction is quantifiable. Suppose person A has $1 million and person B has only $10,000. Suppose also that the MU of an additional $10,000 for person A is 10 utils, but the MU of an additional $10,000 for person B is 500 utils. If the law of diminishing marginal utility holds, then this assumption seems to make sense. That is, the relatively poor individual values the additional $10,000 much more than the wealthy individual. It seems that we might be able to make a case for a redistribution of wealth from person A to person B strictly on efficiency grounds. Society will enjoy a net gain of about 490 utils due to the transfer.[5]

On the other hand, if we assume ordinal utility, then all we can assert is that both individuals will prefer the larger sum ($1,010,000 for person A and $20,000 for person B) to the smaller sum ($1,000,000 for person A and $10,000 for person B). We cannot say how much more each person prefers the larger sum to the smaller sum. As a result, it is impossible to conclude that a redistribution of money will lead to a net welfare gain. With ordinal utility, **interpersonal utility comparisons** are impossible. That is, who is to say that the individual with $1 million will value an additional $10,000 less than the individual with only $10,000? Because we lack a common yardstick for the measurement of satisfaction, we cannot justify the redistribution of wealth on efficiency grounds. If any argument is to be made for the redistribution of wealth, it must appeal to our desire for greater equality or some standard of fairness. For those wishing to limit government redistributions of wealth and income, the ordinal utility assumption is a powerful one.

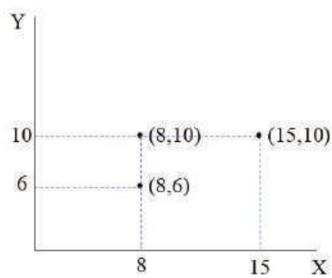
The Assumptions of the Modern Theory of Utility Maximization

Just as in the case of the traditional theory, it is necessary to specify the conditions in which the consumer makes decisions with the goal of maximizing utility. It is again assumed that the consumer is choosing between two goods, X and Y. The consumer also possesses a fixed money income (M) and faces a constant price for good X (P_X) and a constant price for good Y (P_Y). The consumer also has given individual preferences. The only major difference between the two theories is that the consumer's preferences must be represented differently because in the modern theory, utility is not quantifiable.

Budgets, Preferences, and Utility Maximization

To graphically represent the modern theory of utility maximization, we will be working in a two-dimensional space called the **commodity space**. An example is shown in Figure 6.5 where the quantities of goods Y and X are measured on the vertical and horizontal axes, respectively.

Figure 6.5: The Commodity Space



Each point in the commodity space represents a **commodity bundle**, or a combination of the two goods. For example, the ordered pair (8, 6) represents a commodity bundle that includes 8 units of good X and 6 units of good Y.

Our purpose is to determine which of the affordable bundles in the commodity space maximizes the consumer's utility. We will approach this problem in three parts. First, we will determine the set of all affordable bundles in the commodity space for a given income level and prices. Second, we will explain how preferences may be represented when utility is ordinal using an analytical device designed for this purpose known as indifference curves. Finally, we will bring together the consumer's budget with her preferences to determine the utility maximizing commodity bundle.

In terms of the consumer's budget, we know that the money income and the prices of the two goods are fixed. Let's first see if we can determine all the commodity bundles that require the consumer to spend all her income. In other words, let's assume the following condition:

$$Money\ Income = Consumer\ Expenditure$$

If the consumer spends the entire money income, then no saving exists. This same condition can be written symbolically in the following way:

$$\overline{M} = \overline{P_X}X + \overline{P_Y}Y$$

The overbars in the equation indicate that the prices and income are constant. Only the quantities of goods X and Y are variables in the equation because only they can be influenced by the consumer's choices. To graph this equation in the commodity space, we need to solve for the variable on the vertical axis. If we solve for Y, we obtain the following result (omitting the overbars):

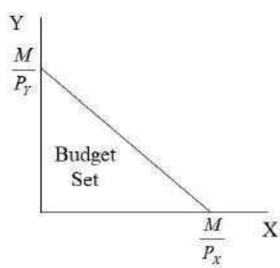
$$Y = \frac{M}{P_Y} - \frac{P_X}{P_Y}X$$

This equation is in slope-intercept form and so it is now possible to graph this equation as a straight line in the commodity space. The vertical intercept is constant and is equal to M/P_Y. The slope is also constant and is equal to $-P_X/P_Y$. The horizontal intercept can also be determined by setting Y equal to 0 and solving for X as follows:

$$0 = \frac{M}{P_Y} - \frac{P_X}{P_Y} X \Rightarrow X = \frac{M}{P_X}$$

All these results are captured in Figure 6.6:

Figure 6.6: The Budget Set



In economic terms, the vertical and horizontal intercepts refer to the maximum quantities of Y and X, respectively, that can be purchased if the consumer spends the entire money income on one or the other good. Each of the other points on the **budget line** represents a commodity bundle that requires the consumer to spend the entire money income. Bundles below the budget line are affordable but do not require the expenditure of the entire income. For example, if a consumer began at a bundle on the budget line and reduced the amount of Y by a small amount, then some income would be left over. Hence, that bundle would be affordable but would not require the expenditure of the entire income. All the bundles on the budget line and those below the budget line are affordable and are therefore considered part of the consumer's **budget set**. Any bundle above the budget line is not affordable because the consumer would need more income than she possesses to make that purchase.

The slope of the budget line is also important and can be written as follows:

$$\frac{\Delta Y}{\Delta X} = -\frac{P_X}{P_Y}$$

The slope of the budget line represents the ability of the consumer to trade off one good for another in the marketplace at the current prices. That is, if the consumer would like to purchase one more unit of X, she will have to give up $-P_X/P_Y$ units of Y. The reader might be thrown off by the fact that P_X is in the numerator and P_Y is in the denominator. It should be remembered that the units in which price is measured, however, are $ per unit. Therefore, P_X is measured in $/X and P_Y is measured in $/Y. The division yields Y/X, which are the correct units attached to the slope of the budget line.

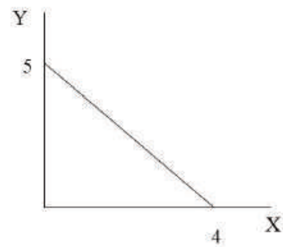
To make this discussion a bit more concrete, let's consider a numerical example. Suppose M = $20, P_X = $5, and P_Y = $4. We can write the budget equation and then solve for Y as follows:

$$20 = 5X + 4Y$$

$$Y = 5 - \frac{5}{4}X$$

This budget equation may be graphed as in Figure 6.7.

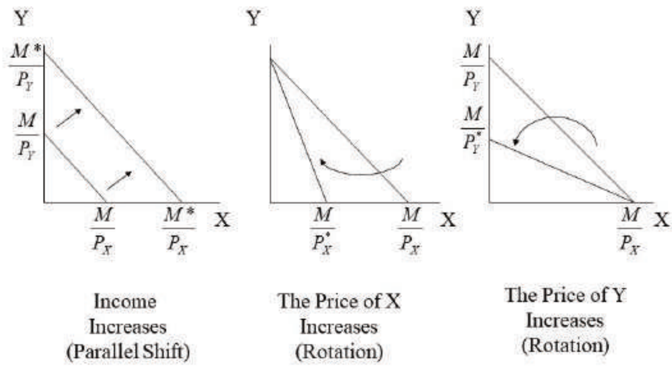
Figure 6.7: An Example of a Budget Line



The reader should notice that the intercepts may be obtained quickly since $M/P_X = 4$ and $M/P_Y = 5$. The slope may also be determined quickly because $-P_X/P_Y = -5/4 = -1.25$. That is, if the consumer wishes to increase the consumption of X by one unit, then he must give up 1.25 units of Y.

It is possible that the consumer's money income or the prices of the goods may change. If such changes occur, then the budget set will expand or contract as the budget line moves. For example, Figure 6.8 shows what will happen if changes occur in M, P_X, or P_Y.

Figure 6.8: Changes in the Budget Set



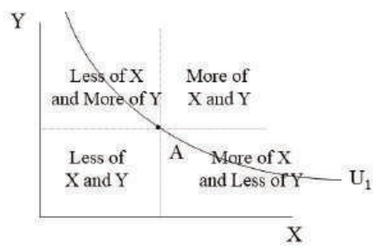
As the figure shows, a rise in M will increase both intercepts proportionately, leading to an outward, parallel *shift* of the budget line. More income thus makes more bundles affordable. Similarly, a rise in the price of X leads to a *rotation* inwards of the budget line because the vertical intercept is unaffected. In this case, the budget set contracts because the higher price of X makes fewer bundles affordable. Finally, a rise in the price of Y also makes the budget line rotate inwards because the horizontal intercept is unaffected. The contraction of the budget set indicates that fewer bundles are affordable. The reader should work through what will happen when the consumer's income or the prices of the goods fall.

We now turn to the representation of preferences within the modern theory of utility maximization. The key analytical tool for this purpose is a device called an **indifference curve**. It is defined as follows:

- An indifference curve through commodity bundle A is a line drawn through all the commodity bundles that the consumer considers to be equally satisfying in relation to commodity bundle A.

An example of an indifference curve is given in Figure 6.9.

Figure 6.9: An Indifference Curve

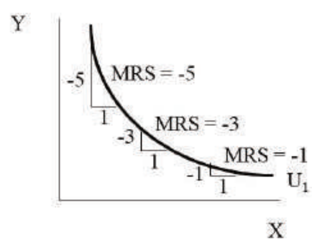


All the bundles along indifference curve, U_1, are equally preferred to all other bundles on that same indifference curve. It is also possible to identify regions that have more of both goods relative to bundle A, less of both goods relative to bundle A, more Y and less X relative to bundle A, and less Y and more X relative to bundle A.

Our next task is to explain why the indifference curve through bundle A slopes downward as well as why it is bowed away from the origin (as opposed to a straight line or having some other shape). The downward slope indicates that the consumer is willing to trade off some of good Y for more of good X and vice versa. For example, if the consumer begins at bundle A, and she loses some of good Y, then the only way to restore her to the U_1 level of satisfaction is to increase her consumption of good X. Therefore, the indifference curve slopes downward.

The fact that the indifference curve is bowed towards the origin indicates something significant about the willingness of the consumer to trade off one good for another. To clarify this point, we will refer to the slope of the indifference curve as the **marginal rate of substitution** (MRS). Figure 6.10 shows clearly that the slope is very steep at low levels of X and becomes very flat at high levels of X.

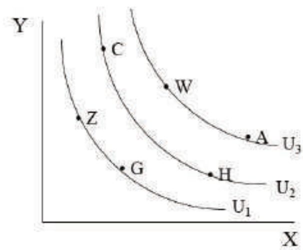
Figure 6.10: The Marginal Rate of Substitution (MRS)



As the figure shows, when the MRS is – 5 (approximately) the consumer is willing to sacrifice 5 units of Y for one more unit of X. When the MRS is – 3 (approximately) the consumer is willing to sacrifice 3 units of Y for 1 more unit of X. Finally, when the MRS is – 1 the consumer is willing to sacrifice only 1 unit of Y for 1 more unit of X. Clearly, as the consumer consumes more X, the amount of Y he is willing to sacrifice to obtain another unit of X becomes smaller and smaller. In other words, X is becoming less valuable to the consumer as more of it is consumed. Although this result sounds a great deal like the law of diminishing marginal utility, it was obtained without the use of cardinal (quantifiable) utility. The only reason that we know the consumer values additional units of X less as more of it is consumed is that he is willing to give up less and less Y for those additional units. **Diminishing marginal rate of substitution** is the phrase we can use to refer to this tendency of the MRS to fall (in absolute value) as more X is consumed.

Finally, it needs to be emphasized that the consumer is assumed to have a *complete* set of preferences. In other words, every bundle can be ranked against every other bundle in the commodity space. Therefore, an infinite number of indifference curves must exist with one passing through every possible bundle. A series of indifference curves is called an **indifference map**. Because an infinite number exists, it is not possible to show all the indifference curves in the commodity space. Figure 6.11 shows one example of an indifference map.

Figure 6.11: An Indifference Map

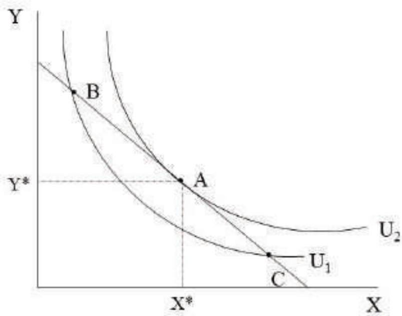


In the figure, a few commodity bundles have been labeled. Is it possible to identify the consumer's ranking of these bundles? It is so long as we assume that more of both goods is always preferred to less of both goods. If we make this assumption, then it follows that any commodity bundle on a higher indifference curve must be preferred to any bundle on a lower indifference curve. In this case, W and A are equally preferred. W and A are strictly preferred to C and H, which are equally preferred to one another. Finally, Z and G are equally preferred to one another, and they are less preferred than C and H. We are free to state that W and A give the consumer the most utility, but since we are using an ordinal notion of utility, this statement only means that they are the most preferred bundles in the ranking.

We have finally reached the point where we can determine the consumer's utility maximizing choice given her income and the prices she faces. Figure 6.12

shows a consumer's budget line and two of her indifference curves.

Figure 6.12: The Utility Maximizing Choice



We want to determine what the utility maximizing commodity bundle is. We can begin by narrowing down the possibilities very quickly. The bundle that maximizes utility must be on the budget line. If it is not, then the consumer will have income left over. Because only the consumption of goods generates utility, the consumer cannot be maximizing utility unless she is consuming on the budget line.

Suppose we consider a bundle on the budget line such as bundle B. Because this bundle is on the budget line, we know that all income is being spent. Does it maximize utility though? It does not. The reason is that the indifference curve through bundle B (U_1) is lower than the indifference curve through bundle A (U_2). Therefore, the consumer prefers bundle A to bundle B. Because bundle A is affordable, the consumer should buy less Y and more X until she reaches bundle A. Similarly, if the consumer is consuming bundle C, then it is possible to reach the higher indifference curve by reducing her consumption of X and increasing her consumption of Y until she reaches bundle A where she consumes X* units of X and Y* units of Y.

At bundle A, the consumer is maximizing utility because all income is spent, and it is not possible to reach a higher indifference curve by choosing a bundle within the budget set. At that bundle, the budget line is tangent to the highest indifference curve passing through the budget set. Mathematically, the slope of the budget line must equal the slope of the indifference curve. We can express this condition symbolically as:

$$-\frac{P_X}{P_Y} = MRS$$

This utility maximizing condition may be written another way as well. Let's first recall that total utility remains constant as we move along an indifference curve. That is, ΔTU = 0. When moving along the indifference curve, we can assume that the consumer reduces her consumption of good Y (by ΔY) and increases her consumption of good X (by ΔX). The change in total utility can now be divided up into the loss of utility from the reduced consumption of Y *plus* the gain in utility from the increased consumption of X. That is:

$$\Delta TU = \underbrace{MU_Y \cdot \Delta Y}_{Loss} + \underbrace{MU_X \cdot \Delta X}_{Gain} = 0$$

.

If we solve the above equation for ΔY/ΔX, we obtain the following result:
$$\frac{\Delta Y}{\Delta X} = -\frac{MU_X}{MU_Y}$$

Even though utility is not measurable in the modern framework, the utils cancel out in the fraction and so the MRS is a ratio of the change in Y to the change in X. We include the marginal utilities because this new manner of expressing the MRS allows us to write the utility maximizing condition in the following way:
$$-\frac{P_X}{P_Y} = -\frac{MU_X}{MU_Y}$$

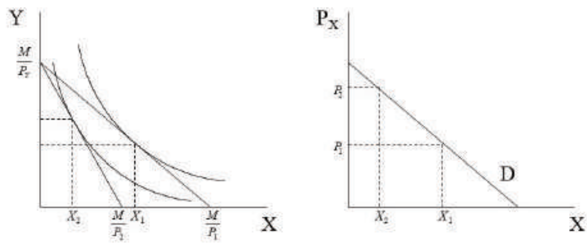
By rearranging the terms and eliminating the negative signs, we obtain the same result that we obtained in the traditional theory of utility maximization.
$$\frac{MU_X}{P_X} = \frac{MU_Y}{P_Y}$$

In other words, the marginal utilities per dollar spent must be the same for both goods for utility to be maximized. This result was a remarkable outcome for neoclassical economists. They obtained precisely the same result that they had obtained with the traditional theory, but they did so without the less realistic assumption of cardinal utility. Other things the same, the theory with the simplest assumptions became the chosen theory. Ockam's Razor had been successfully applied.

The Modern Derivation of the Individual Demand Curve

The modern theory of utility maximization also allows us to derive the downward sloping individual demand curve. Figure 6.13 provides the graphical analysis by which the demand curve is derived within a utility maximizing framework.

Figure 6.13: The Modern Derivation of the Individual Demand Curve



In Figure 6.13, the price of good X rises from P_1 to P_2, causing a rotation inward of the budget line. The consumer moves from one utility maximizing point to another. In this case, he reduces the quantity demanded of X from X_1 units to X_2 units. The result is a movement up the individual demand curve. Because the consumer moves to a lower indifference curve, it is apparent that the consumer's level of satisfaction has diminished as he moves up the demand curve. This conclusion makes sense because the consumer now faces a higher price for good X and nothing else has changed about his situation. This analysis could easily be reversed with a decline in the price of good X and a rise in the quantity demanded of good X.

Theories of Endogenous Preferences

Throughout this discussion of consumer behavior, we have been treating preferences as given or fixed in accordance with the neoclassical entry point. When preferences are treated as given or fixed, they are said to be **exogenous**. That is, they are determined by external factors, and their determination is not explained within the model.

Heterodox critiques of neoclassical theory often focus on the assumption of exogenously determined preferences. Many critics argue that preferences should be treated as **endogenously** determined because often the variables that do change in neoclassical models can be argued to have a direct effect on preferences. One reason neoclassical economists have generally avoided making preferences endogenous in their models is that it becomes very difficult mathematically to incorporate this assumption into formal models. Furthermore, the assumption of endogenous preferences makes it impossible to derive the demand curve using the modern approach to utility maximization.

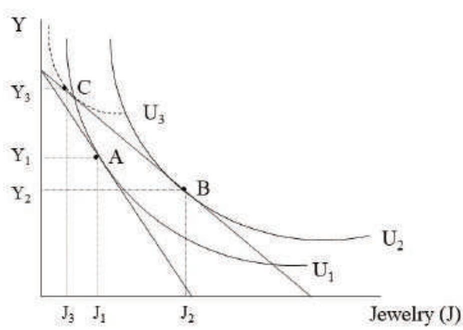
To understand how the assumption of endogenous preferences interferes with the derivation of the demand curve, consider the case of **conspicuous consumption** first analyzed in detail by the great American social theorist Thorstein Veblen (1857-1929). According to Veblen, the upper classes would place their acts of consumption and leisure on display for all to observe. Examples include mansions, yachts, luxury vehicles, expensive art works, jewelry, fine clothes, and servants. According to the description of two historians of economic thought:

> For the wealthy, the more useless and expensive a thing

> was, the more it was prized as an article of conspicuous consumption. Anything that was useful and affordable to common people was thought to be vulgar and tasteless.[6]

The term **Veblen Good** is sometimes used to refer to goods for which the quantity demanded rises as the price *increases*. A wealthy person thus wishes to purchase more of an item as its price rises rather than less because of the message it conveys to others about that person's ability to buy expensive items. What this possibility implies is that the consumer's preferences change because the price rises. Therefore, preferences are endogenous. To illustrate this point, consider Figure 6.14, which shows how a consumer's utility maximizing choice of jewelry changes as the price falls.

Figure 6.14: A Fall in the Price of a Veblen Good



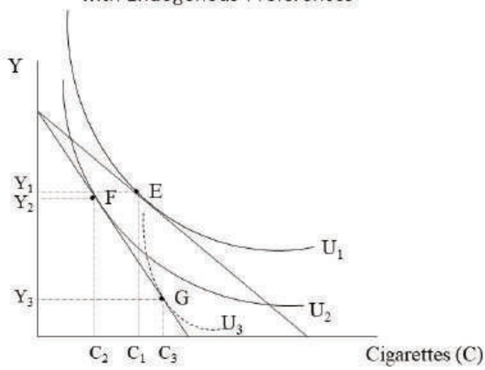
According to Figure 6.14, as the price of jewelry falls, the budget line rotates outward as usual. If preferences are given, then the quantity of jewelry that the consumer chooses increases from J_1 to J_2, and the consumer's utility level rises from U_1 to U_2. In this case, however, preferences are endogenously determined. Because the price of jewelry has fallen, this wealthy consumer's preferences change such that she prefers to purchase less jewelry rather than more. The indifference curves *move* such that the utility maximizing choice of jewelry is J_3 at the lower price, and the utility level is U_3.[7] Of course, less wealthy consumers might not experience any change in preferences when the price of jewelry falls, and they will purchase more.

One might think that the lower price of jewelry and the smaller quantity purchased suggests an upward sloping demand curve. That is, price and quantity demanded appear to be positively related in this case. The problem is that the demand curve shows the relationship between price and quantity demanded when everything else is held constant, including preferences. Because the preferences change when the price changes, it is impossible to derive the individual demand curve for a Veblen good.

The Veblen Good is not the only example where preferences are likely to be endogenously determined. Consider the case of cigarettes. According to data available from the Centers for Disease Control and Prevention (CDC), in 2006 cigarette companies spent $12.4 billion on advertising and promotional expenses in the United States. The CDC reports that specific populations have been the targets of these advertising campaigns, including young people, women, and racial/ethnic communities.[8]

Let's consider an example in which a high school student experiences an increase in the price of cigarettes per pack. As usual, a rise in the price of cigarettes causes a rotation inward of the budget line as shown in Figure 6.15.

Figure 6.15: A Rise in the Price of Cigarettes with Endogenous Preferences



If preferences are exogenously determined, then the student will reduce his consumption of cigarettes from C_1 to C_2 and his utility will decline from U_1 to U_2. If preferences are endogenously determined, on the other hand, then it is possible that the rise in price will lead to greater profits for the tobacco companies. This increase in profits may be used to more aggressively advertise tobacco products to young people. The advertising campaign might change the preferences of the student so much that the student ends up purchasing more cigarettes even though the price rose. In Figure 6.15, the indifference curves move such that the final consumption of cigarettes is C_3 and the student's utility level is U_3.

Even though the price of cigarettes rose, the student bought more. This positive relationship might lead one to expect an upward sloping demand curve. As in the case of the Veblen Good, it is impossible here to derive the individual demand curve because preferences have not remained constant. These two examples demonstrate how serious the complications are that arise from the assumption that preferences are endogenous. Without the ability to derive an individual demand curve, the supply and demand explanation of market price completely collapses.

Following the Economic News[9]

In 2012 a fire at a Bangladeshi factory killed over 112 workers. The company that had subcontracted orders to the factory was a Walmart supplier. The drive for profit has led many suppliers to participate in a race to the bottom in developing nations to force prices down. Research indicates that some Americans are willing to pay more for clothes from factories that don't exploit workers. American consumers can try to educate themselves about factory conditions, but the problem is that many times this information is not immediately available to the consumer.

According to Richard Locke, the Deputy Dean of MIT's Sloan School of Management, consumers need to break the cycle by purchasing less of this cheap clothing from retail stores. Here we find another instance in which endogenous preferences may be relevant. If the price of a clothing item declines, a consumer may interpret the price reduction as an indication of worse working conditions for workers. A consumer wishing to break the cycle of falling prices and worker exploitation might choose to buy less of this cheap apparel rather than more. That is, the price reduction has set into motion events that have led to a change in the consumer's preferences. Of course, purchasing less of this clothing may cause workers in developing nations to lose their jobs and income, as unpleasant as those changes may be. Unfortunately, no easy solution to this problem exists.

Summary of Key Points

1. The positive theory of utility maximization developed by neoclassical economists in the 1870s has its roots in the normative philosophy of utilitarianism.
2. The law of diminishing marginal utility states that the extra satisfaction, from the consumption of an additional unit of a good, declines as the amount consumed rises.
3. In the traditional theory of utility maximization, the consumer will spend all income such that the marginal utilities per dollar spent are the same for each good.
4. The paradox of value can be resolved using either the classical labor theory of value or the marginal utility concept.
5. Consumers' surplus and producers' surplus assume a constant marginal utility of money and one that is the same for all buyers and sellers.
6. Ockam's Razor states that the simplest theory is the best theory, other things equal, and for that reason the ordinal utility assumption is superior to the cardinal utility assumption.
7. Changes in price cause rotations of the budget line whereas changes in income cause parallel shifts of the budget line.

8. Indifference curves slope downwards because consumers are willing to make tradeoffs, and they are bowed towards the origin because consumers experience diminishing marginal satisfaction as they consume more of a good.
9. Utility maximization in the modern theory requires that the slope of the budget line equal the marginal rate of substitution.
10. When preferences are endogenous, the individual demand curve cannot be derived because a change in price causes preferences to change.

List of Key Terms

Utilitarianism

Marginal utility

Utility

Utils

Total utility

Marginal utility

Law of diminishing marginal utility

Paradox

Cardinal utility

Ordinal utility

Interpersonal utility comparisons

Commodity space

Commodity bundle

Budget line

Budget set

Indifference curve

Marginal rate of substitution (MRS)

Diminishing marginal rate of substitution

Indifference map

Exogenous preferences

Endogenous preferences

Conspicuous consumption

Veblen Good

Problems for Review

1. Suppose Steve is currently spending all his money income on pencils and erasers. The price of a pencil is \$.75 and the price of an eraser is \$1.50. If the marginal utility that Steve derives from the last pencil purchased is 5 utils and the marginal utility of the last eraser purchased is 12 utils, then will Steve buy more pencils, fewer pencils, or the same number of pencils? Why?

2. Assume that the price of good A is \$1 per unit and the price of good B is \$2 per unit. Complete the rest of the table.

Problems for Review: Problem 2

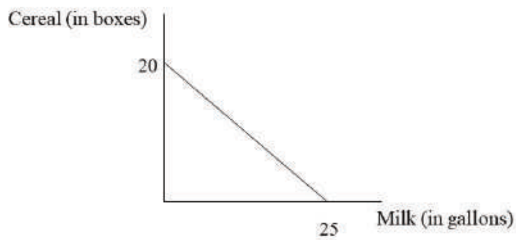
Unit	TU_A	MU_A	MU_A/P_A	TU_B	MU_B	MU_B/P_B
First		12		20		
Second	22				18	
Third		8		54		
Fourth	36				14	
Fifth		4		80		

3. Using the information from problem 2, assume that John has $14 of money income to spend on the two goods. How many units of each good will John purchase?

4. Suppose Bill has an income of $26 and faces constant prices for yogurt and pudding. Yogurt is $4 per pack and pudding is $5 per pack. Derive Bill's budget equation and draw it on a graph. Label both axes and both intercepts. Place yogurt on the vertical axis.

5. Consider the budget constraint shown in Figure 6.16. Suppose the price of cereal rises from $2.75 to $5.50 per box. What is the consumer's income if the line represents the consumer's initial budget line? Draw the new budget line on the same graph and label any new intercepts. Can you determine the price of milk (per gallon)?

Figure 6.16: Problem 5



6. Suppose the price of cigarettes (per pack) falls. How might the quantity purchased by a high school student be affected if her preferences are exogenously determined? How might her preferences be affected if her preferences are endogenously determined? Explain.

Notes

1. Samuelson and Nordhaus (2001), p. 85, encourage readers to "resist the idea that utility is a psychological function or feeling that can be observed or measured." Instead, they regard it as a "scientific construct."
2. It is possible to construct an example in which a consumer has some money left over if that amount is not sufficient to pay the price of an additional unit of either good.
3. The classical economists also recognized that the scarcity or

abundance of a good would influence its price, but we can think of that characteristic as influencing the labor requirement. For example, if water is plentiful in a region, then the labor time required for its production will be lower.

4. Neoclassical textbooks frequently discuss this paradox, but they typically do not acknowledge that the classical economists possessed an internally consistent solution to the paradox. I would like to acknowledge Prof. David Ruccio, whose lecture on this topic in an introductory economics class at the University of Notre Dame in the early 2000s influenced my presentation of this topic. I served as a teaching assistant for Prof. Ruccio at the time.
5. The loss of utility for person A will probably be slightly more than 10 utils due to diminishing marginal utility.
6. Hunt and Lautzenheiser (2011), p. 338.
7. The reader should note that indifference curves with given preferences never cross. The only reason they appear to cross in this graph is because two different sets of indifference curves are being placed on the same graph.
8. Centers for Disease Control and Prevention, http://www.cdc.gov/tobacco/data_statistics/fact_sheets/tobacco_industry/marketing/index.htm
9. Bajaj, Vikas. "Before You Buy That T-Shirt." *The New York Times*. May 18, 2013.

CHAPTER 7

THEORIES OF PRODUCTION TECHNOLOGY AND COST OF PRODUCTION

Goals and Objectives:

In this chapter, we will do the following:

1. *Describe* short run production technology in the neoclassical tradition
2. *Define* the law of diminishing marginal returns to a single input
3. *Explain* how short run production technology generates short run production cost
4. *Analyze* different kinds of long run production technologies in the neoclassical tradition
5. *Derive* long run production cost from long run production technology
6. *Demonstrate* the relationship between short run and long run production cost
7. *Investigate* Schumpeterian and Marxian theories of production technology and cost

Short Run Production Technology in the Neoclassical Tradition

In the neoclassical tradition, **production technology** is defined very broadly to include all of the accumulated knowledge and technical skills that are used in the production of goods and services. As we will see, the unidirectional logic that is so central to neoclassical thinking applies in this case in that technology is used to determine the cost structure that firms face.

We begin with a distinction between two periods. The **short run** is a period in which at least one factor of production is fixed. If a pizzeria uses only capital (K) and labor (L), for example, we might expect labor to be a variable input whereas capital is its fixed input. That is, it may be possible for the firm to hire anywhere from 0 to a very large (unspecified) number of workers. At the same time, the restaurant itself, which includes its dining area, kitchen, and other facilities, is fixed in size. If the firm is not able to build a new restaurant or to add on to the existing restaurant, the firm is operating in the short run. If it takes 6 months to construct a new restaurant, then the short run in this specific industry is 6 months. The **long run** then refers to a period during which all factors of production are variable. In this example, the long run refers to a period that is longer than 6 months because both capital and labor are variable.

It is worth noting that the specific periods corresponding to the short run and the long run vary by industry. It may take two years to construct a new production plant in one industry and only six months to construct a new factory in a second industry. The amount of time simply depends on how long it takes for all inputs to become variable. It should also be emphasized that production technology remains fixed in both the short run and the long run. The knowledge and skills related to production do not change in this analysis, which is a key entry point of neoclassical theory. Also, consistent with the neoclassical approach is the assumption of a fixed capital endowment in the short run, which is another key entry point.

We will now continue with the example of a pizzeria in the short run. To represent the firm's short run production technology, we will introduce a few key short run product concepts. The first concept is **total product** (TP). The total product refers to the total physical output produced during a given period, in this case a specific number of pizzas. In earlier chapters, we referred to total output as Q, which has the same meaning. Therefore, we can write the following:

$$TP = Q$$

An additional product concept that is critical for representing production technology in neoclassical theory is the concept of **marginal product** (MP). When an additional worker is hired, that additional worker increases total pizza production by a specific amount. This amount is the marginal product of that worker. For example, if the third worker increases total pizza production from 11 to 14 pizzas, then the marginal product of the third worker is 3 pizzas. Formally, we write the marginal product in the following way:

$$MP = \frac{\Delta TP}{\Delta L} = \frac{\Delta Q}{\Delta L}$$

Finally, the concept of **average product** (AP) is important as well. In this example, the average product of labor refers to the average number of pizzas produced per worker. For example, if the total product is 30 pizzas and 10 workers are employed, then the average product is 3 pizzas per worker. It is important to understand the difference between average product and marginal product. Whereas marginal product only considers the addition to the total product of the last worker hired, average product spreads out the entire product evenly over the number of employed workers. The average product may be written as:

$$AP = \frac{TP}{L} = \frac{Q}{L}$$

Table 7.1 provides an example with some hypothetical data for our pizzeria operating in the short run. Let's say the period we are considering is one business day.

Table 7.1: Production Technology for a Pizzeria Operating in the Short Run

Labor (L)	Total Product (TP)	Marginal Product (MP)	Average Product (AP)
0	0	–	–
1	8	8	8
2	20	12	10
3	36	16	12
4	48	12	12
5	56	8	11.2
6	60	4	10
7	62	2	8.86
8	62	0	7.75
9	60	-2	6.67

Using the information in the table, the MP when 3 workers are employed is calculated in the following way:

$$MP = \frac{\Delta TP}{\Delta L} = \frac{36-20}{3-2} = \frac{16}{1} = 16\ pizzas\ per\ worker$$

Similarly, the AP of the third worker is calculated as follows:

$$AP = \frac{TP}{L} = \frac{36}{3} = 12\ pizzas\ per\ worker$$

We are now able to consider the patterns that emerge from these calculations. The MP appears to rise until it reaches its peak at L = 3 workers. Next it declines, eventually reaching 0, and then becomes negative. The range over which MP rises is called the range of **increasing returns to labor**. The question arises as to why MP rises as more workers are hired at these low employment levels. The answer is **specialization** and the **division of labor**. Just think about all of the tasks that are involved in the pizza business. If a single worker is hired, then that worker needs to make the dough, roll the dough, toss the dough, make the sauce, spread the sauce, add the toppings, bake the pizza, cut the pizza into slices, take customer orders, serve the pizza to customers, bus the tables, wash the dishes, operate the register, clean the kitchen, clean the dining area, and clean the restrooms! The worker needs to accomplish other tasks as well, including ordering more ingredients, maintaining financial records, and paying taxes. Clearly, if a second worker is hired, then the tasks can be divided up, allowing for a greater degree of specialization. These gains from specialization imply that the second worker increases the TP by 12 pizzas, which is more than the increase that resulted from the hiring of the first worker. When three workers are hired, the tasks can be further divided so that one worker can work in the kitchen, a second worker can handle customer service in the dining area, and a third worker can work the cash register and maintain the books. As a result, the MP continues to rise. The reader should be aware that the additional workers hired are not more skilled or more hardworking. In fact, we are assuming the opposite. That is, each worker is equally skilled and hardworking. These gains in MP are solely the result of the benefits of specialization and the division of labor. The AP rises as well for this same reason.

Once we move beyond the point where three workers are hired, we begin to see a decline in the MP of labor. The range over which MP falls is called the range of **diminishing returns to labor**. Why does this decline in MP occur? Imagine that the kitchen is very small at our restaurant and the fourth worker is asked to work in the kitchen with the cook already present there. The worker contributes to production and so TP rises. At the same time, the presence of the additional worker in the kitchen makes it somewhat more difficult for the original cook to work efficiently. The two cooks keep running into one another. At times, they bump elbows as one rolls the dough and the other spreads the sauce. Even though the fourth worker adds to production, TP only rises by 12 pizzas rather than 16 pizzas as a result of this "bumping elbows problem." The more formal phrase that is used to refer to this phenomenon is the **law of diminishing returns** to labor. It should be noted that AP falls for the same reason.

The law of diminishing returns may be expressed in two different ways. It may be expressed as:

1. a decline in the marginal product of labor as employment increases
2. an increase in the total product of labor at a diminishing rate

The reader should refer to Table 7.1 to observe these two patterns. It should be noted that this problem of bumping elbows eventually becomes so serious that the TP begins to fall. In other words, the MP becomes negative as additional workers reduce the total production of pizzas. The range of negative MP is referred to as the range of **negative returns to labor**. The underlying reason for this pattern of diminishing returns and then negative returns to labor is that the capital input (i.e., the size of the restaurant, and the kitchen, specifically) is fixed in the short run. The same pattern would eventually result if we were to add additional workers in the dining area and in the finance area. Due to the importance of the fixed capital input in generating this result, the law of diminishing returns is a specifically short run phenomenon.

It really should be emphasized how extraordinary the claim is that neoclassical economists are making when they assert that the law of diminishing returns holds for production technologies in the short run. They are claiming that every single short run production process exhibits this same pattern. It does not matter whether we are discussing the production of commercial aircraft, high-tech computers, cheese pizzas, automobiles, or haircuts. All these production processes have this basic feature of their production technologies in common. It is a fundamental aspect of our economic existence.

It is also possible to represent these patterns graphically. The TP curve that corresponds to the data in Table 7.1 is represented in Figure 7.1. The MP and AP curves corresponding to these data are represented in Figure 7.2.

Figure 7.1: The Total Product Curve

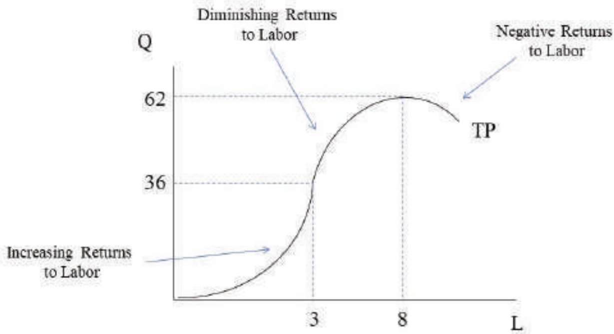
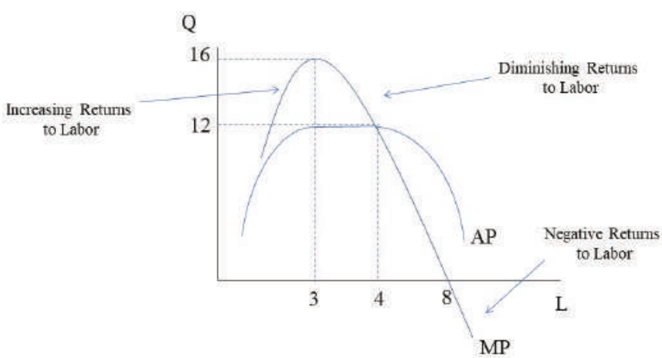


Figure 7.2: The Marginal Product and Average Product Curves



It is important to notice that MP and AP intersect in Figure 7.2 at L = 4 when both the AP and the MP are equal to 12 pizzas per worker. It turns out that MP always intersects AP at the peak of AP. Why? It helps to consider a non-economic example first. Suppose that the total weight of everyone in a classroom of 30 people is 4,200 lbs. The average weight of a person in the room is then equal to 140 lbs. (= 4,200 lbs. / 30 people). Next let's consider what would happen if we bring in a marginal person (person A) from outside the room. Person A is a football player who weighs 300 lbs. Notice that the total weight in the room is now 4,500 lbs. and the number of people in the room is 31 people. In that case, the new average weight will be approximately 145 lbs. (= 4,500 lbs. / 31 people). Because the marginal weight exceeded the average weight, the average weight rose. It follows that if the marginal contribution exceeds the average, the average will rise. In the case of MP and AP, if the MP exceeds the AP, the AP will rise.

If instead we had brought in a marginal person (person B) with a weight of 100 lbs., then the average weight of a person in the room would be approximately equal to 139 lbs. (= 4,300 lbs. / 31 people). Because the marginal weight is below the average weight in this example, the average weight falls. It follows that if the marginal contribution is below the average, the average will fall. In the case of MP and AP, if the MP is below the AP, the AP will fall.

In Figure 7.2, the reader can see that at low employment levels, the MP exceeds the AP causing the AP to rise. At higher employment levels, the MP falls below the AP causing the AP to fall. It follows that the MP must intersect the AP at the peak of the AP curve. To the left of the peak, the higher MP pulls the AP up and to the right of the peak, the lower MP pulls the AP down.

Short Run Production Cost in the Neoclassical Tradition

Our next task is to explain the cost structure that firms face in the short run. We will differentiate between two types of short run cost. First, **total variable cost** (TVC) refers to the total monetary cost of the variable input, which in this case is labor. If we assume a constant wage rate (w), measured in terms of dollars per worker, then TVC may be calculated as follows:
$$TVC = wL$$

Because L can be measured in terms of the number of workers employed during this period, TVC is measured in dollars of labor cost incurred during this period. TVC can change in the short run because L is a variable input.

The other major type of cost is **total fixed cost** (TFC). TFC refers to the total monetary cost of the fixed input, which in this case is capital. If we assume a constant price of capital (P_K), measured in terms of dollars per unit of capital, then TFC may be calculated as follows:
$$TFC = P_K K$$

Because K is measured in terms of the number of units of capital employed during this period, TFC is measured in dollars of capital cost incurred during this period. With both K and P_K fixed, the TFC is also fixed in the short run.

The firm's **total cost** (TC) is simply the sum of TVC and TFC. It can be expressed as:
$$TC = TVC + TFC = wL + P_K K$$

TC is measured in terms of dollars of cost incurred during this period.

For example, we can select some hypothetical cost data for a pizzeria operating in the short run as shown in Table 7.2.[1]

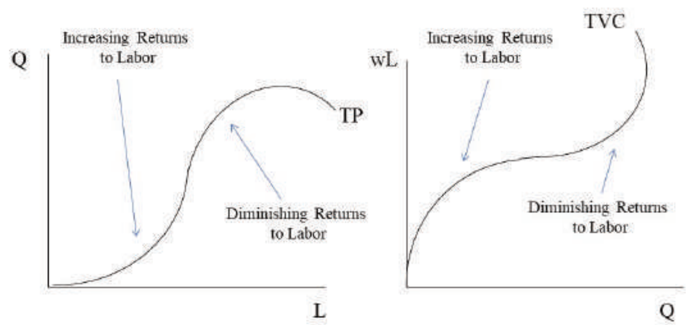
Table 7.2: Total Cost Information for a Pizzeria Operating in the Short Run

Total Product (TP)	Total Fixed Cost (TFC)	Total Variable Cost (TVC)	Total Cost (TC)
0	$150	$0	$150
10	150	80	230
20	150	150	300
30	150	210	360
40	150	290	440
50	150	380	530
60	150	480	630
70	150	590	740
80	150	710	860
90	150	840	990

If the numbers in Table 7.2 correspond to one business day, then the capital cost or TFC (i.e., the rent for the building) is fixed for all levels of output at $150 per day. The labor cost or TVC, on the other hand, rises with output because increased employment of the variable input is what makes possible the rise in output. Finally, total cost rises because the TVC is rising even though the TFC remains fixed. It should also be noted that TVC is $0 when 0 units of output are produced because no workers are hired, but TFC is still $150 because the rent must be paid regardless of whether pizzas are produced.

It is possible to say more about this pattern of short run production cost. In fact, we can use the short run production technology concepts already developed to derive a specific pattern of short run production cost. It is easiest to understand this derivation by considering the similarities between the TP curve and the TVC curve as shown in Figure 7.3.

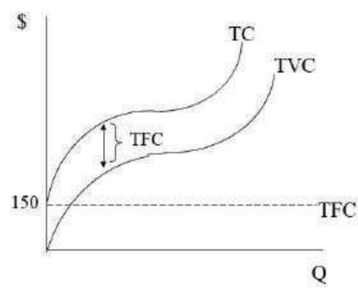
Figure 7.3: The Derivation of the Total Variable Cost Curve



First consider the axes in each of the graphs in Figure 7.3. The axes are essentially reversed in the two graphs. The only other difference is the multiplication of L by the w in the TVC graph. If w is $1 per worker, then this difference disappears. Even if the wage is some other value, the only result will be a stretching of the graph, but its basic shape will remain unchanged. Now consider the shapes themselves. The reader might imagine lifting the TP curve off the graph and twisting it around until it can be placed directly on top of the TVC curve. What this exercise implies is that the TVC curve possesses its shape for precisely the same reason that the TP curve possesses its shape. That is, at low output levels, TVC rises slowly due to the specialization and division of labor. At higher output levels, TVC rises more quickly due to diminishing returns to labor as it becomes necessary to hire more and more labor to increase production by given amounts.

The next step is to consider the shapes of the TFC and TC curves as shown in Figure 7.4.

Figure 7.4: The Total Cost Curves



Because the TFC is fixed at $150 for all levels of output, the TFC curve is simply a horizontal line. The TC curve, however, has the same shape as the TVC curve because TC only changes due to changes in TVC (since TFC is fixed). In addition, the TC curve is shifted upwards by the amount of the TFC. As a result, the distance between the TVC and TC curves is always equal to the TFC. Because the TFC is reflected in the distance between these two curves, it is not necessary to include the TFC curve when we wish to represent the three total cost measures on a single graph. For that reason, the TFC curve has been included here as a dotted line. Also, because all three concepts are measured in dollars, the "$" symbol has been placed on the vertical axis.

Next, we need to introduce several unit cost concepts. Often managers of firms think in terms of production cost per unit of output produced, particularly when they are considering whether to increase or decrease production by a small quantity. We will define four key unit cost concepts, and then we will consider the pattern these unit cost measures exhibit as output changes in the short run.

The first unit cost concept that we will define is **average variable cost** (AVC). The AVC is the variable cost per unit of output produced. If labor is the variable input, then the AVC is per unit labor cost. If total wages amount to $1000 and 100 units of output are produced, then the AVC is $10 per unit (= $1000/100 units). It is defined in the following way:

$AVC = \frac{TVC}{Q}$

Another unit cost concept that we will define is **average fixed cost** (AFC). The AFC is the fixed cost per unit of output produced. If capital is the fixed input, then the AFC is the per unit capital cost. If total capital cost amounts to $4500 and 100 units of output are produced, then the AFC is $45 per unit (= $4500/100 units). It is defined in the following way:

$AFC = \frac{TFC}{Q}$

A third unit cost concept that we will define is **average total cost** (ATC). The ATC is the total cost per unit of output produced. It is simply the sum of the AVC and the AFC. If the AVC is $10 per unit and the AFC is $45 per unit, then the ATC is $55 per unit (= 10+45). Alternatively, if total cost is $5500 and total output is

100 units, then ATC is $55 per unit (= $5500/100 units). It is defined as follows:

$$ATC = \frac{TC}{Q} = \frac{TVC+TFC}{Q} = \frac{TVC}{Q} + \frac{TFC}{Q} = AVC + AFC$$

The final unit cost concept that we will define is **marginal cost** (MC). The MC is the additional total cost per additional unit of output produced. For example, if total cost rises by $200 and total output rises by 50 units, then MC is $4 per unit. Furthermore, because total cost can only rise due to a rise in variable cost (since TFC is fixed), the MC can also be defined as the additional variable cost per additional unit of output produced. It is thus defined as follows:

$$MC = \frac{\Delta TC}{\Delta Q} = \frac{\Delta TVC}{\Delta Q}$$

Using the total cost information presented in Table 7.2, we can now calculate all these additional unit cost measures for our pizzeria operating in the short run. Table 7.3 combines the total cost information from Table 7.2 with the corresponding unit cost calculations.

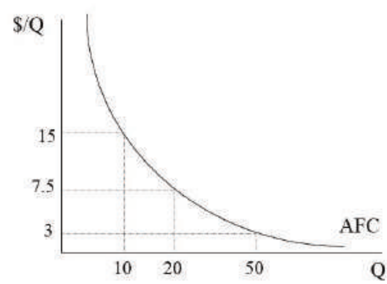
Table 7.3: Total Cost and Unit Cost Information for a Pizzeria Operating in the Short Run

TP	TFC	TVC	TC	AFC	AVC	ATC	MC
0	$150	$0	$150	--	--	--	--
10	150	80	230	15.00	8.00	23.00	8.00
20	150	150	300	7.50	7.50	15.00	7.00
30	150	210	360	5.00	7.00	12.00	6.00
40	150	290	440	3.75	7.25	11.00	8.00
50	150	380	530	3.00	7.60	10.60	9.00
60	150	480	630	2.50	8.00	10.50	10.00
70	150	590	740	2.14	8.43	10.57	11.00
80	150	710	860	1.88	8.88	10.75	12.00
90	150	840	990	1.67	9.33	11.00	13.00

Clear patterns emerge for each of the unit cost concepts.

The AFC, for example, declines continuously as total product rises. This reduction in AFC occurs because the fixed cost is spread over a larger and larger quantity of output. Businesspeople refer to this phenomenon of a falling AFC as “spreading one’s overhead.” Figure 7.5 shows a graph of the AFC curve.

Figure 7.5: The Average Fixed Cost Curve

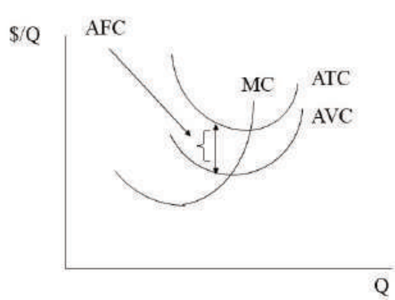


As Figure 7.5 indicates, the AFC curve possesses horizontal and vertical asymptotes. That is, at very low output levels, the AFC soars because the denominator approaches zero while the numerator remains fixed (equal to TFC). As output rises, on the other hand, AFC declines due to the spreading of overhead. That is, output becomes increasingly large while the numerator remains fixed. The curve will never cross the horizontal axis though because some fixed cost will always exist no matter how thinly it is spread across the existing output. It should also be noted that $/Q has been placed on the vertical axis because the unit cost concepts are measured in dollars per unit of output.

In addition, Table 7.3 shows that the AVC, ATC, and MC measures all decrease at low output levels and then rise at higher output levels. Because they are derived from the total cost measures, the reductions in AVC and MC can be attributed to specialization and division of labor. The ATC falls at low output levels for two reasons. As the sum of AVC and AFC, it falls due to specialization and division of labor (falling AVC) and due to the spreading of overhead (falling AFC). All three measures begin to rise at higher output levels due to diminishing returns to labor.

As Figure 7.6 shows, the AVC, ATC, and MC curves are all U-shaped.

Figure 7.6: Relationships among the Unit-Cost Curves



Furthermore, because the ATC is equal to the sum of AFC and AVC, it follows that AFC = ATC − AVC. Therefore, the difference between the ATC and AVC curves will always equal the AFC. As a result, we do not need to include the AFC curve when we wish to graphically represent the four unit-cost measures. For that reason, the AFC curve has been omitted in Figure 7.6. It should also be noticed that the AVC and ATC curves grow closer together as Q rises. The reason is that AFC is declining due to the spreading of overhead.

An additional point that must be made is that MC intersects AVC and ATC at the minimum points on the AVC and ATC curves. The reason is the same as that used to explain why MP intersects AP at the maximum AP. That is, when MC is below the AVC, the AVC falls. When the MC is above the AVC, the AVC rises. Therefore, the MC intersects the AVC at the minimum AVC. The same argument can be used to explain why MC intersects ATC at the minimum ATC. In terms of Table 7.3, we can see that MC changes from being below AVC to being greater than AVC between 30 and 40 units of output. The intersection of the two curves, therefore, occurs within that range. Similarly, MC changes from being below ATC to being greater than ATC between 60 and 70 units of output. The intersection of the MC and ATC curves must occur within that range as well.

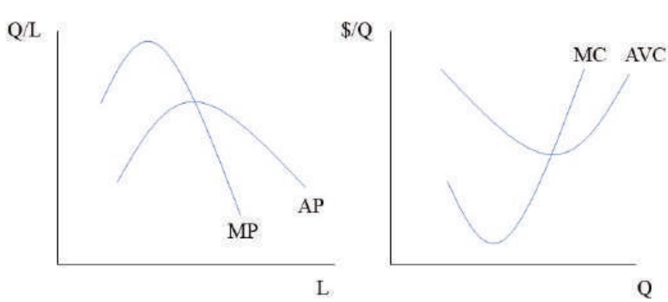
Finally, the relationship between the product curves and cost curves should be emphasized once more. Because production technology determines production cost in neoclassical theory, we should be able to identify some connection between the product concepts and the unit cost concepts. It turns out that the MP and AP curves are the mirror image of the MC and AVC curves,

respectively. If we look at the definitions, this implication leaps out at us.
$MP = \frac{\Delta TP}{\Delta L} = \frac{\Delta Q}{\Delta L}\ and\ MC = \frac{\Delta TVC}{\Delta Q} = \frac{\Delta wL}{\Delta Q} = \frac{w\Delta L}{\Delta Q}$

Clearly, if ΔQ/ΔL rises, then ΔL/ΔQ falls. That is, if MP rises, then MC falls. The opposite also holds that if MP falls, then MC rises. We can make the same argument about the relationship between AP and AVC. That is:
$AP = \frac{TP}{L} = \frac{Q}{L}\ and\ AVC = \frac{TVC}{Q} = \frac{wL}{Q}$

It should be clear that if Q/L rises, then L/Q falls. Hence, a rise in AP coincides with a fall in AVC. The opposite also holds that if AP falls, then AVC rises. Figure 7.7 shows these curves side-by-side to emphasize the relationships between them.

Figure 7.7: The Relationship Between the Product Curves and the Unit Cost Curves



Finally, we have been assuming the ceteris paribus condition throughout this analysis. That is, we have assumed that the wage rate, the price of capital, the quantity of capital, and the production technology have remained fixed. We should briefly consider what will happen to the unit cost curves if any of these factors changes. For example, a rise in wages will cause the AVC, ATC, and MC curves to shift upwards. The reason is that wages determine TVC, which influences the position of each of these curves. If wages rise, the firm faces higher unit costs. Second, a rise in the price of capital or in the quantity of capital employed will cause an upward shift of the ATC curve but will leave the AVC and MC curves unchanged. The reason is that these factors only affect the TFC, which influences the AFC but not the AVC or MC. Hence, the gap between the ATC and AVC curves will rise as AFC grows but otherwise nothing will change. Finally, a technological improvement will cause a reduction in the AVC and MC (and thus the ATC). Therefore, a downward shift of the AVC, MC, and ATC curves will occur. It is left to the reader to consider the consequences of a reduction in wages, the price of capital, or the quantity of capital employed. A loss of technological knowledge is also theoretically possible and should be considered.

A Short Run Technology and Cost Example Using Microsoft Excel

It is helpful to use a spreadsheet software program such as Microsoft Excel when analyzing the short run production technology and cost structure facing a firm. The reason is that it allows one to avoid performing all the calculations by hand, and so it saves the analyst a great deal of work. Let's consider a simple example in which an analyst only has a small amount of information about a firm. Suppose the analyst only knows that the wage is \$5 per hour and the price of capital (P_K) is \$2 per unit. That is, w = 5 and P_K = 2. Suppose that the only other information the analyst possesses relates to total product and the capital stock as shown in Table 7.4.

Table 7.4: A Short Run Technology and Cost Example Using MS Excel (Incomplete)

	A	B	C	D	E	F	G	H	I	J	K	L
1	L	K	TP	AP	MP	TVC	TFC	TC	AFC	AVC	ATC	MC
2	0	50	0	=C2/A2	=(C2-C1)/(A2-A1)	=5*A2	=2*B2	=F2+G2	=G2/C2	=F2/C2	=H2/C2	=(H2-H1)/(C2-C1)
3	1		5									
4	2		15									
5	3		30									
6	4		50									
7	5		75									
8	6		105									
9	7		140									
10	8		180									
11	9		225									
12	10		275									
13	11		320									
14	12		360									
15	13		395									
16	14		425									
17	15		450									
18	16		470									
19	17		485									
20	18		495									
21	19		500									

In MS Excel, the columns are labeled with capital letters and the rows are numbered. The reader may be surprised to learn that the entire table can be completed with only the information that has been provided. To complete the entire table by hand would require 199 separate calculations! Fortunately, MS Excel allows us to use the definitions of each measure and then use cell references to complete the entire table in a small fraction of the time it would take to complete the table by hand.

Because the capital stock is fixed in the short run at all output levels, we can simply copy and paste that quantity of capital to every cell in column B. To carry out this operation quickly, it is only necessary to place the cursor over the lower right corner of the cell B2 until the cursor turns into a little black cross (+). The analyst may then drag the cursor all the way down the column until it is complete.

The other columns are a little more difficult to complete because it is necessary to insert a formula to ensure an accurate calculation. Whenever entering a formula in MS Excel, it is necessary to first include an "=" sign. For example, because the definition of AP is TP/L, the AP is calculated as =C2/A2 for L = 0. As with the capital stock, this formula can be dragged down to complete the rest of the cells in the column. The rest of the columns are completed in a similar fashion. It should be noted that the "*" symbol, which is used in the calculation of TVC and TFC, is used for the multiplication of terms in Excel.

Table 7.5 shows the complete table after all calculations have been performed.

Table 7.5: A Short Run Technology and Cost Example Using MS Excel (Complete)

	A	B	C	D	E	F	G	H	I	J	K	L
1	L	K	TP	AP	MP	TVC	TFC	TC	AFC	AVC	ATC	MC
2	0	50	0	#DIV/0!	#####	0	100	100	#DIV/0!	#DIV/0!	#DIV/0!	######
3	1	50	5	5.00	5	5	100	105	20.0000	1.0000	21.0000	1.0000
4	2	50	15	7.50	10	10	100	110	6.6667	0.6667	7.3333	0.5000
5	3	50	30	10.00	15	15	100	115	3.3333	0.5000	3.8333	0.3333
6	4	50	50	12.50	20	20	100	120	2.0000	0.4000	2.4000	0.2500
7	5	50	75	15.00	25	25	100	125	1.3333	0.3333	1.6667	0.2000
8	6	50	105	17.50	30	30	100	130	0.9524	0.2857	1.2381	0.1667
9	7	50	140	20.00	35	35	100	135	0.7143	0.2500	0.9643	0.1429
10	8	50	180	22.50	40	40	100	140	0.5556	0.2222	0.7778	0.1250
11	9	50	225	25.00	45	45	100	145	0.4444	0.2000	0.6444	0.1111
12	10	50	275	27.50	50	50	100	150	0.3636	0.1818	0.5455	0.1000
13	11	50	320	29.09	45	55	100	155	0.3125	0.1719	0.4844	0.1111
14	12	50	360	30.00	40	60	100	160	0.2778	0.1667	0.4444	0.1250
15	13	50	395	30.38	35	65	100	165	0.2532	0.1646	0.4177	0.1429
16	14	50	425	30.36	30	70	100	170	0.2353	0.1647	0.4000	0.1667
17	15	50	450	30.00	25	75	100	175	0.2222	0.1667	0.3889	0.2000
18	16	50	470	29.38	20	80	100	180	0.2128	0.1702	0.3830	0.2500
19	17	50	485	28.53	15	85	100	185	0.2062	0.1753	0.3814	0.3333
20	18	50	495	27.50	10	90	100	190	0.2020	0.1818	0.3838	0.5000
21	19	50	500	26.32	5	95	100	195	0.2000	0.1900	0.3900	1.0000

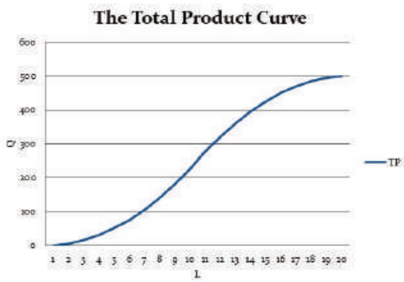
A few points worth noticing include the fact that diminishing returns to labor first appear at L = 11 when MP shows a decline for the first time. Additionally, it is also at this employment level that MC begins to rise for the first time, which demonstrates the notion that the MP and MC are mirror images of one another. Finally, it is possible to locate the points of intersection of MC with AVC and ATC. MC intersects AVC between employment levels 13 and 14, whereas MC intersects ATC between employment levels 17 and 18. The reader might notice that MP intersects AP between employment levels 13 and 14, a fact which also reinforces the point that the MP and AP curves are mirror images of the MC and AVC curves.

It is also possible to generate graphs in Excel of the product and cost curves. To illustrate how these graphs may be created, we will create a graph of the TP curve and a separate graph containing the AP and MP curves. To create a graph, the analyst should select the following:

1. On the Insert Tab, choose "Line."
2. From the options available, choose "Line" again.
3. A chart box should appear. Right click it and choose "Select Data."
4. Next click "Add" and in the new box type "TP" for Series Name.
5. Where it says, "Series Values," click the symbol to the right and then drag in the data for TP from the Excel table beginning with L = 0. Then click OK and click OK once more.
6. To add horizontal and vertical axis labels, you can use the Design and Layout Tabs.

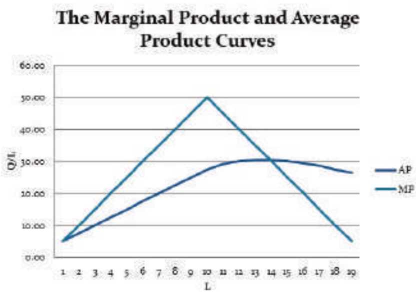
Figure 7.8 shows the TP curve as generated in MS Excel.

Figure 7.8: The Total Product Curve as Generated in MS Excel



The same procedure may be used to create the graph of the AP and MP curves. The only difference is that you need to return to step 4 to "Add" the second curve after the first has been created. Also, because the values are undefined for L = 0, it is best to omit these calculations when dragging in the data for these two curves. Figure 7.9 shows the AP and MP curves as generated in Excel.

Figure 7.9: The AP and MP Curves as Generated in MS Excel



Long Run Production Technology in the Neoclassical Tradition

We now turn to production technology in the long run when all inputs are variable. In the case of the pizzeria, it is now technically possible for the firm to construct additional restaurants in addition to hiring additional workers. We will discuss three types of long run production technologies before we explain how these long run technologies determine long run production costs.

The first type of long run production technology is a technology that exhibits **constant returns to scale** (CRS). A CRS technology exists when a fixed percentage change in all inputs leads to the same percentage change in total output. For example, suppose that the amounts of capital and labor double (100% increases in each). In that case, the amount of output produced also exactly doubles, assuming a CRS technology. Using the pizzeria example, if 3 workers in one restaurant produce 10 pizzas in an hour, then 6 workers in two restaurants will produce 20 pizzas in an hour, assuming a CRS technology. The assumption of a CRS technology should seem reasonable enough. A firm should be able to perfectly duplicate its smaller scale operation when it doubles its inputs and so output should exactly double.

A second type of long run production technology is a technology that exhibits **increasing returns to scale** (IRS). An IRS technology exists when a fixed percentage change in all inputs leads to a greater percentage change in total output. For example, suppose that the amounts of capital and labor double (100% increases in each). In that case, the amount of output produced *more than* doubles, assuming an IRS technology. It may seem strange, given what was previously explained about the likely consequences of duplication, that output could more than double. It would be relatively easy to explain this possibility with an *impure* example. That is, we might argue that the *quantities* of labor and capital have exactly doubled but the *qualities* of the newly acquired capital and labor are superior to what was previously used, causing output to more than double.

It is possible, however, to provide a *pure* example of an IRS technology or one in which a doubling of the inputs leads to a greater than doubling of output even when the additional inputs are of the same quality as that which was previously used. Such cases arise when the production processes have certain geometric characteristics that make IRS technologies possible.[2] For example, consider an oil pipeline that transports oil over a long distance. To keep the example simple, we will focus on the opening at one end of the pipeline, which is assumed to form a perfect circle. The circumference of the circle (i.e., the distance around it) is directly related to the amount of material that is used to construct the pipeline. That is, a larger circumference implies a larger amount of material that is required to produce the pipeline. The circumference (C) is given by the following formula where r is the radius of the circle:

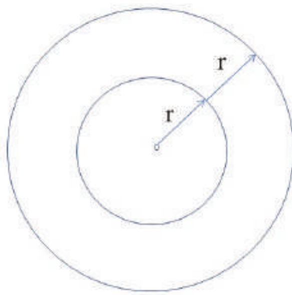
$$C = 2\pi r$$

The area of the circle, by contrast, is directly related to the amount of oil that can pass through the pipeline. That is, a larger opening implies that a larger amount of oil can pass through the pipeline. The area of the circle is given by the following formula:

$$A = \pi r^2$$

Now consider what happens when we double the amount of material used in the construction of the pipeline. This doubling of the amount of material makes possible a doubling of the circumference of the circle. In Figure 7.10, the inner circle represents the original pipeline opening and the outer circle represents the new pipeline opening that uses twice as much material.

Figure 7.10: The Scale Expansion of An Oil Pipeline Opening



According to the figure, the doubling of the circumference doubles the radius of the circle. As a result, we can calculate the circumference of the larger opening (C') as:

$C' = 2\pi(2r) = 4\pi r = 2C$

In other words, the doubling of the radius has doubled the circumference, which indicates that the amount of material inputs has doubled. What effect does this doubling of the radius have on the area of the new opening? We can calculate the area of the larger opening (A') as follows:

$A' = \pi(2r)^2 = 4\pi r^2 = 4A$

In other words, the doubling of the radius leads to a quadrupling of the area of the opening. That is, the material inputs in the production of the pipeline have doubled and yet the output that passes through the pipeline has more than doubled. It appears we have a pure example of an increasing returns to scale (IRS) technology.

The third and final type of long run production technology is a technology that exhibits **decreasing returns to scale** (DRS). A DRS technology exists when a fixed percentage change in all inputs leads to a smaller percentage change in total output. For example, suppose that the amounts of capital and labor double (100% increases in each). In that case, the amount of output produced *less than* doubles, assuming a DRS technology. It is difficult to provide *pure* examples of DRS technologies, although it is relatively easy to provide any number of *impure* examples. That is, we might argue that the *quantities* of labor and capital have exactly doubled but the *qualities* of the newly acquired capital and labor are inferior to what was previously used, causing output to less than double. More will be said about these issues when we discuss long run production cost.

Table 7.6 provides a summary of the three types of long run production technologies.

Table 7.6: An Overview of Long Run Production Technologies

Stage	Technology	Labor (L)	Capital (K)	Output (Q)
1	----	50	1	1000
2A	CRS	100	2	2000
2B	IRS	100	2	3000
2C	DRS	100	2	1500

In each of the three cases (2A, 2B, and 2C), the quantities of capital and labor exactly double, but the output responds differently in each case. In the case of CRS, output exactly doubles. In the case of IRS, output triples. Finally, in the case of DRS, output only rises by 50%.

Long Run Production Cost in the Neoclassical Tradition

We are now able to draw some conclusions about long production cost based on our discussion of long run production technology. As in the short run, production technology determines the pattern of production cost that the firm faces. The most important long run cost concept that we will explore is **long run average total cost** (LRATC). As in the short run, average total cost in the long run is simply total cost per unit of output. The only difference is that LRATC may change due to a change in capital cost or due to a change in labor cost. It is defined as follows:

$$LRATC = \frac{TC}{Q} = \frac{wL+P_KK}{Q}$$

Let's assume that the factor prices are constant in the long run such that w = 1 and P_K = 2. We can now see clearly what the relationship is between the three types of long run production technology and LRATC. Table 7.7 shows the calculations for TC and LRATC for each of the three cases considered previously in Table 7.6.

Table 7.7: The Relationship Between Long Run Technology and LRATC

Stage	Technology	Labor (L)	Capital (K)	Output (Q)	Total Cost (TC)	LRATC
1		50	1	1000	52	$.05
2A	CRS	100	2	2000	104	$.05
2B	IRS	100	2	3000	104	$.03
2C	DRS	100	2	1500	104	$.07

What this example shows clearly is that total cost exactly doubles in all three scenarios because the inputs have exactly doubled. It is the changes in output that differ in all three scenarios, which causes the LRATC to be different in the three cases. Specifically, we see that an increase in the scale of production leaves LRATC unchanged at $0.05 per unit in the case of a CRS technology. Because TC and Q grow at the same rate, the ratio of the one to the other does not change. By contrast, an increase in the scale of production causes the LRATC to fall from \$0.05 to \$0.03 per unit in the case of an IRS technology. The reason is that TC grows more slowly than Q and so the ratio of TC to Q declines. Finally, an increase in the scale of production causes the LRATC to rise from \$0.05 to \$0.07 per unit in the case of a DRS technology. The reason is that TC grows more quickly than Q and so the ratio of TC to Q rises.

The calculations from this example can be summarized as follows:

$$For\ CRS,\ LRATC = \frac{w(2L)+P_K(2K)}{2Q} = \frac{2TC}{2Q} = \frac{TC}{Q} \Rightarrow \overline{LRATC}$$

$$For\ IRS,\ LRATC = \frac{w(2L)+P_K(2K)}{3Q} = \frac{2TC}{3Q} = \frac{2}{3}\frac{TC}{Q} \Rightarrow LRATC \downarrow$$

$$For\ DRS,\ LRATC = \frac{w(2L)+P_K(2K)}{1.5Q} = \frac{2TC}{1.5Q} = \frac{4}{3}\frac{TC}{Q} \Rightarrow LRATC \uparrow$$

Special phrases are used to describe changes in LRATC as the scale of production changes:

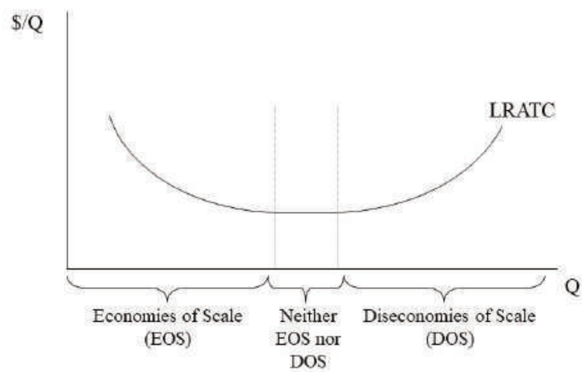
1. **Economies of scale** (EOS) refer to reductions in per unit cost as output rises.
2. **Diseconomies of scale** (DOS) refer to increases in per unit cost as output rises.

When per unit cost does not change with an expansion in the scale of production, the firm is said to experience neither EOS nor DOS.

Next, we should consider which of these cost patterns is most likely for an individual firm. Neoclassical economists argue that firms typically pass through all three phases as the scale of production grows. Specifically, firms experience scale economies at relatively low output levels, then they experience neither

EOS nor DOS, and finally, they experience diseconomies of scale at relatively high output levels. Graphically, this pattern gives rise to a U-shaped LRATC curve like the one shown in Figure 7.11.

Figure 7.11: The Long-Run Average Total Cost Curve



Why do neoclassical economists assert that this pattern of unit cost is a typical one facing firms? They give several reasons for expecting scale economies at relatively low output levels.[3] The first is labor and managerial specialization. Unlike in the short run, these types of specialization require that all inputs be variable. For example, imagine a firm that carries out both production and distribution activities in a single facility. As it expands in the long run and increases its capital stock, it might be able to carry out production in one facility and distribution in another facility. The gains from specialization may be considerable. As a result, TC grows more slowly than Q and LRATC falls.

Another reason for scale economies at relatively low output levels may be large startup costs. For example, in the automobile industry a huge outlay of financial capital is required to even begin operations. As more automobiles are produced, this initial outlay can be spread over a larger number of units and LRATC falls. Large startup costs in the long run operate in a similar manner to TFC in the short run.

As firms become larger, they also frequently enjoy efficiencies that stem from expanded knowledge of the production process. That is, they simply become more efficient producers as the kinks in the production process are worked out. As a result, TC rises less quickly than Q and LRATC falls. Other factors that contribute to scale economies have more to do with input prices. For example, as firms expand they can take advantage of bulk discounts that suppliers offer. Another possibility is that they can borrow funds more cheaply in the stock and bond markets than smaller firms, which must rely on bank financing. As a result, output grows more quickly than cost and LRATC falls.

On the other end of the spectrum are scale diseconomies, which can arise for a variety of reasons. First, corporate bureaucracies may become excessively large as the scale of production expands. If highly paid top executives are hired to prepare a lot of paper reports, the result may be a considerable rise in TC without much expansion of Q. In that case, LRATC rises. Communication problems are another potential source of DOS in large organizations. If managers have a difficult time coordinating production due to the large size, then further expansion may raise TC more than Q.

The alienation that workers experience in the workplace in very large organizations may be another problem. Poor work performance may result when workers feel disconnected from one another and from their work. The sense of alienation may become so great that Q rises slowly even as TC increases with the purchase of additional inputs. The additional cost of hiring supervisors to monitor worker performance in these situations may also add to cost. Although Marxists tend to emphasize worker alienation to a much greater degree than neoclassical economists, neoclassical economists sometimes recognize that this factor contributes to DOS.[4]

Finally, input prices may also be involved in creating DOS. For example, as firms expand to very large sizes, they may begin to push up input prices due to their increased demand for inputs. Similarly, if their large sizes necessitate transporting goods to more distant markets, then TC may rise more quickly than output.[5]

The Relationship between Short Run and Long Run Average Total Cost

The LRATC curve is sometimes described as the lower envelope of all **short run average total cost** (SRATC) curves. To understand the reason, consider Figure 7.12.

Figure 7.12: The Relationship Between the Long Run and Short Run Average Total Cost Curves

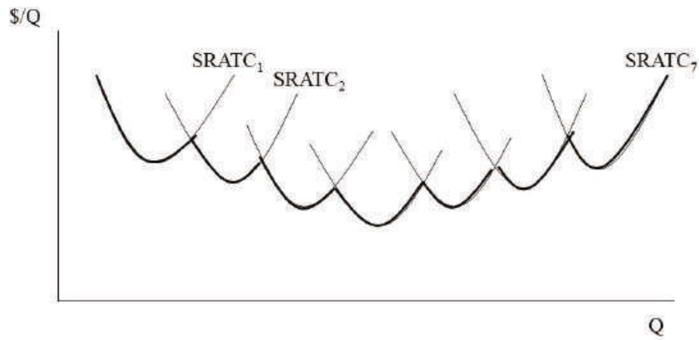
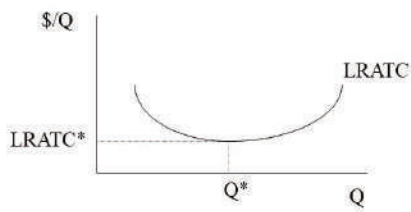


Figure 7.12 shows a series of SRATC curves, each corresponding to a different production plant size. As production increases in the short run for a firm operating with plant size 1, unit cost declines due to labor specialization within the plant. Eventually, diminishing returns begin to set in as output expands due to the fixed amount of capital, and unit cost rises. In the long run, however, the firm has the option of building a larger production plant, which allows the firm to move to plant size 2. As a result, unit cost begins to decline again due to specialization within the larger plant. Eventually, however, diminishing returns result, and unit cost begins to rise. The firm can expand in the long run by building an even larger production plant, and the same process repeats itself. Due to this analysis, we can trace out the lower envelope of the SRATC curves to obtain the LRATC curve.

Minimum Efficient Scale

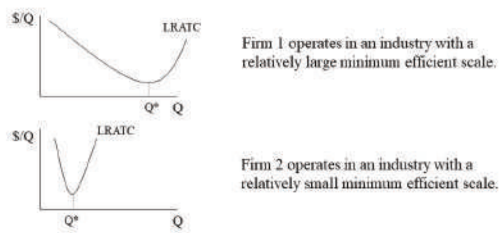
A final concept that deserves attention is what neoclassical economists call **minimum efficient scale** (MES). The MES refers to the lowest output level at which LRATC is minimized. Figure 7.13 shows the MES to be Q* because it is the lowest output level that minimizes LRATC.

Figure 7.13: Minimum Efficient Scale (MES)



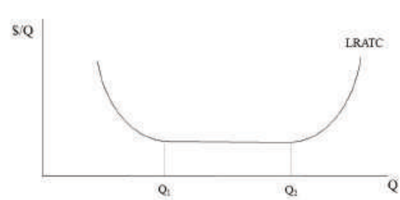
The MES varies by industry. For example, the MES has been estimated to be 10 million tons per year in the oil industry, 9 million tons per year in the steel industry, and at least 1 million barrels per year in the beer industry.[6] The MES in an industry may, therefore, be a large amount of output or a small amount of output. For example, Figure 7.14 contains examples of two LRATC curves. Each represents the LRATC of a firm in a different industry. Firm 1's LRATC curve shows that the MES is very large due to large economies of scale in this industry. Only large firms in this industry will be able to minimize unit cost. Firm 2's LRATC curve suggests a very small MES due to significant diseconomies of scale in this industry. Only small firms will be able to minimize unit cost in this industry.

Figure 7.14: Variation in LRATC by Industry



Furthermore, a firm may have multiple production levels that are consistent with minimum LRATC. For example, the firm's LRATC curve shown in Figure 7.15 indicates that Q_1 is the minimum efficient scale, but Q_2 represents the **maximum efficient scale**.

Figure 7.15: A LRATC Curve with Many Output Levels that Minimize Unit Cost



That is, Q_2 is the largest output level that still manages to keep unit cost at its minimum. In this industry, small firms, large firms, and medium-sized firms can operate at minimum unit cost.

The Schumpeterian Approach to Technological Change

The neoclassical theory of production technology and cost of production that we have been considering treats technology as completely static. Almost no emphasis is placed on technological change. Granted, neoclassical economists have more to say about technological change in macroeconomics, but in microeconomics it plays essentially no active role at all. For this reason, many heterodox economists challenge the neoclassical treatment of production technology.

One important example of an economist with a different approach to production technology is the famous economist Joseph Schumpeter, who is best known for his book, *The Theory of Economic Development*, first published in 1913. According to Schumpeter, technological change is the driving force in economies that creates economic growth, unstable and uneven as it is. Meghnad Desai provides a nice overview of Schumpeter's theory:

> Innovations often clustered together and made for an Industrial Revolution. But they came only periodically and discontinuously, in waves. One wave of innovations would break the mould of the old stationary economy. The pioneers would borrow money and launch their innovations. They would enjoy a temporary monopoly, and make immense profits (if, that is, they were successful, since there are also failures in this life). These innovations would launch the economy on a growth path with an upswing that might last twenty to twenty-five years. But

> imitators would soon come in and erode the entrepreneurs' excess profits. Output would multiply, but prices would start falling and profits would go down. A downswing would set in until another wave of innovations hit the economy.[7]

According to Schumpeter, innovation would lead to long periods of growth and progress followed by periods of falling prices and sluggish growth until a new series of innovations revived the growth of capitalism. When innovation does occur, it generates growth in the innovating sectors but leads to "creative destruction" for the sectors made obsolete by the new technology. Clearly, Schumpeter's theory shifts our focus to macroeconomics and away from microeconomics, but his theory also reveals just how static and unhelpful the microeconomic theory of production technology is when it comes to explaining technological progress.

A Marxian Approach to Technological Change

Much can be said regarding Marx's thoughts about the role of technological change in capitalist economies and its ability to revolutionize the means of production. Indeed, Marx was a major influence on Schumpeter. When considering the specifically microeconomic aspects of production technology on which neoclassical economists focus, however, we will also find it helpful to contrast with those aspects the competing theory of production technology that is found in Marx's work.

In chapter 3, it was explained that a commodity's value is determined by the socially necessary abstract labor time (SNALT) required for its production. John Eatwell refers to Marx's concept as one of "socially necessary technique."[8] Some producers will devote more labor to production than is socially necessary. Other producers will devote less labor to production than is socially necessary. All producers, however, will receive a value in exchange that is based on the average degree of skill and intensity of labor prevailing in the industry at that point in time. Therefore, producers that use less labor than the social average will enjoy a surplus profit whereas producers that use more labor than the social average will experience a smaller than normal profit. Those producers that use an amount of labor that is exactly equal to the social average will enjoy a normal profit (determined by the amount of surplus value embodied in the commodity).

Because of these discrepancies, the producers will try their best to imitate the techniques of the most efficient producers. The surplus profits will be too tempting. If some of the producers succeed in their imitation of the other producers, then the SNALT embodied in the commodity will decline over time and its value will fall as well.

The value of the commodity is, of course, determined by the average producer. Over time, all producers manage to lower the SNALT that they must perform to produce the commodity as they try to imitate the lowest cost producer and as the lowest cost producer strives for even lower costs. The result is a steady decline in the value of the commodity. In this competitive environment, the surplus profits of the low-cost producer do not rise, even though she is innovating, because the value of the commodity is falling. Similarly, the lower-than-average profits of the high-cost producer persist, even though she is also innovating, because the value of the commodity is falling.

Consider an example that includes a high-cost producer, a low-cost producer, and an average producer. The product of the high-cost producer has an individual, private value of $25 whereas the product of the low-cost producer has an individual, private value of $15. The product of the average producer, however, has an individual, private value of $20. Because the value of the product of the average producer governs the market value, all three producers sell their commodities for $20 each. The low-cost producer thus receives a surplus profit of $5. The average producer does not receive any surplus profit but rather just a normal profit. Finally, the high-cost producer receives $5 less than a normal profit. Over time, technological innovation might cause all the individual, private values to decline by $5. The result will be a reduction in the market value without any change to the surplus profits of each producer. If the degree of technological innovation among the firms differs, however, then the surplus profits of each firm may change.

Following the Economic News

The world is in the midst of a serious food crisis as close to 1 billion people experience severe hunger and roughly an additional billion people are not receiving enough key nutrients. The problem is expected to worsen in the coming decades as the human population continues to increase. With a fixed amount of land for cultivation, the problem of diminishing returns to land makes it difficult to increase food production sufficiently to keep up with global needs. In the past, the problem of diminishing returns to land has not become too serious because of technological advances in agriculture.

Climate change is expected to make the problem worse as the amount of usable farmland diminishes. A UK-government commissioned study has called for urgent action to address the problem of global hunger. Although the report recognizes that some new technologies are still controversial, it suggests that genetic modification, cloning, and nanotechnology may be part of the solution. The report goes much further, however, calling for a "redesign of the whole food system" to put the planet on a sustainable path.[9] The report is correct to do so. The persistence of mass hunger even when large food surpluses exist in some countries is the result of uneven economic development and problems with the redistribution of food to poorer regions.[10] If the global food distribution system responds to market demand rather than human need, we have every reason to expect a continuation of the global food crisis.

Summary of Key Points

1. In the short run, at least one input is fixed whereas in the long run, all inputs are variable.
2. The increasing marginal product of labor is due to specialization and division of labor whereas diminishing returns to labor is due to the inefficiencies resulting from using too much of a variable input with a fixed input.
3. When a marginal contribution exceeds the average, the average will rise. When a marginal contribution falls below the average, the average will fall.
4. Short run unit cost curves fall initially due to the specialization and division of labor, and they rise eventually due to diminishing returns to labor.
5. Constant, increasing, and decreasing returns to

scale are three long run technologies that specify how output changes when all inputs change by a given percentage.
6. Economies and diseconomies of scale describe how unit cost changes as the scale of production changes.
7. Schumpeter asserts that technological innovations are the driving force behind economic development, but this development is highly cyclical.
8. In Marxian theory, competition and technological imitation in an industry will reduce the SNALT required for production and cause the value of a commodity to decline over time.

List of Key Terms

Production technology

Short run

Long run

Total product (TP)

Marginal product (MP)

Average product (AP)

Increasing returns to labor

Specialization and division of labor

Diminishing returns to labor

Law of diminishing returns

Negative returns to labor

Total variable cost (TVC)

Total fixed cost (TFC)

Total cost (TC)

Average variable cost (AVC)

Average fixed cost (AFC)

Average total cost (ATC)

Marginal cost (MC)

Constant returns to scale (CRS)

Increasing returns to scale (IRS)

Decreasing returns to scale (DRS)

Long run average total cost (LRATC)

Economies of scale (EOS)

Diseconomies of scale (DOS)

Short run average total cost (SRATC)

Minimum efficient scale (MES)

Maximum efficient scale

Problems for Review

1.Is it possible for the marginal product of labor to fall while the average product of labor rises? Explain the reason for your answer.

2. The table below represents the short run technology available to a firm. Assume that L is the variable factor of production and K is the fixed factor of production. The price of labor is $6 per unit and the price of capital is $4 per unit. Using a spreadsheet computer program (such as Microsoft Excel), complete the rest of the table.

Problems for Review: Problem 2

L	K	TP	AP	MP	TVC	TFC	TC	AFC	AVC	ATC	MC
0	20	0	–	–				–	–	–	–
1		10									
2		30									
3		60									
4		100									
5		150									
6		210									
7		280									
8		360									
9		450									
10		550									
11		640									
12		720									
13		790									
14		850									
15		900									
16		940									
17		970									
18		990									
19		1000									

3. Using the spreadsheet program you created to fill in the table in problem 2, create the two graphs described below in MS Excel. Remember to label all axes and curves.

- The total product (TP) curve
- The average product (AP) and marginal product (MP) curves (on one graph)

4. In the table from problem 2, what is the level of employment at which diminishing returns begin? What happens to the marginal product of labor and to marginal cost at this employment level?

5. Using the table from problem 2, fill in the blanks. Marginal cost equals minimum AVC between

employment levels ____ and ____. Marginal cost equals minimum ATC between employment levels ____ and ____.

6. Suppose a firm hires only labor and capital and that the price of labor is $8 and the price of capital is $12. In the long run, the firm doubles its workforce from 8 to 16 units and doubles its capital from 12 to 24 units. Its output subsequently increases from 100 to 210 units.

- What sort of technology is this firm using?
- Calculate the firm's long run average total cost before and after it changes its inputs.

Notes

1. These data are not intended to be compatible with the data in Table 7.1.
2. See OpenStax College (2014), p. 170, for a similar application to the chemical industry. Examples of IRS technologies arising from these geometric properties are found in many neoclassical textbooks.
3. These reasons are commonly cited in neoclassical economics textbooks.
4. For example, see McConnell and Brue (2008), p. 393., and Lipsey and Courant, p. 179.
5. See Keat et al. (2013), p. 265.
6. Carlton and Perloff (2000), p. 42.
7. Desai (2004), p. 176-177.
8. Eatwell (1990), p. 342-343.

9. Ghosh, Pallab. BBC News. "Report: Urgent Action Need to Avert Global Hunger." January 24, 2011.
10. Pindyck and Rubinfeld (2013), p. 214.

CHAPTER 8

THEORIES OF INTENSELY COMPETITIVE MARKETS

Goals and Objectives:

In this chapter, we will do the following:

1. *Define* the key revenue and profit concepts that neoclassical economists use
2. *Identify* the defining characteristics of the neoclassical theory of perfect competition
3. *Explain* the two rules of short run profit maximization for perfectly competitive firms
4. *Analyze* five possible cases of short run profit maximization under perfect competition
5. *Derive* the short run supply curve using the rules of short run profit maximization
6. *Investigate* the dynamic adjustment in the case of long run profit maximization
7. *Explore* implications and criticisms of the perfectly competitive model
8. *Examine* Marx's theory of profit rate equalization and the related transformation problem

Neoclassical Revenue and Profit Concepts

In chapter 7, we focused almost exclusively on the factors that determine production cost, or the outlays that are required for a firm to obtain inputs for production. To understand the concept of profit, we must also understand something about the monetary receipts from the sale of the firm's product, also known as total revenue (TR). If a firm only sells a single product then we can write total revenue as we did in chapter 5 where P is the price of the product and Q is the quantity sold. That is:

$$TR = PQ$$

We need to define two other related revenue concepts. The first is **average revenue** (AR), which is the revenue per unit of output sold. It is defined as follows:

$$AR = \frac{TR}{Q}$$

For example, if a firm's TR is $5000 and 1000 units of output are produced, then the AR is $5 per unit.

The other concept that we need to define is **marginal revenue** (MR). Marginal revenue refers to the additional revenue resulting from the sale of an additional unit of output. It is defined as follows:

$$MR = \frac{\Delta TR}{\Delta Q}$$

For example, if a firm's TR rises by $600 and its output increases by 300 units, then the MR is $2 per unit.

We cannot say anything at this stage about the behavior of TR, AR, and MR because we need more specific information about the characteristics of the marketplace. These concepts are developed at greater length in the next section.

First, however, we must consider what neoclassical economists mean by the term "profit," which brings together the revenue and cost concepts that we have been discussing. It turns out that accountants define profit very differently from neoclassical economists, and so it is important to contrast the two definitions of profit. **Accounting profit** refers to the difference between TR and **explicit costs**. Explicit costs include all out-of-pocket costs or monetary costs, such as the payment of wages, rent, and interest. We can write the definition as follows:

$$Accounting\ Profit = TR - Explicit\ Costs$$

By contrast, neoclassical economists argue that accountants do not account for all the costs of production because some costs are not associated with out-of-pocket, monetary payments. Costs that do not carry with them explicit, monetary payments are called **implicit costs**. For example, suppose that the owner of the pizzeria is also the cook. Because she is the owner, she does not pay herself a wage (although she hopes to keep any profit that her firm earns). Because her labor is a resource used in production, it could be used elsewhere for some productive activity. That is, the owner could earn a wage or salary elsewhere and so her labor has an *opportunity cost*. The neoclassical economist, therefore, wishes to include the cost of this labor even though it has no explicit payment associated with it. Similarly, if the owner uses her personal computer to maintain the financial records of the firm, then the cost of this self-owned resource should be included as well even though the business did not directly incur a cost to purchase the computer. Additionally, if the owner invested her own financial capital in the business, then she incurs an opportunity cost in the form of interest and dividend income that would have been earned from other worthy investments. All these implicit costs should be included in any profit calculation, according to neoclassical economists. Neoclassical economists, therefore, define **economic profit** in the following way:

$$Economic\ Profit = TR - Explicit\ Costs - Implicit\ Costs$$

Let's consider an example in which the pizzeria earns $100,000 in revenue in one month. Table 8.1 contains three scenarios. All numerical values are in dollar terms.

Table 8.1: Three Profit/Loss Scenarios for a Pizzeria

		Scenario 1	Scenario 2	Scenario 3
	Total Revenue	100,000	100,000	100,000
Explicit Cost	Wages	70,000	70,000	70,000
	Rent	10,000	10,000	10,000
	Interest	10,000	10,000	10,000
Implicit Cost	Foregone Salary/Interest	10,000	15,000	5,000
	Accounting Profit	10,000	10,000	10,000
	Economic Profit	0	-5000	5000

In each scenario the explicit costs are the same and only the implicit costs differ. Because the explicit costs are the same in all three scenarios, the firm earns $10,000 in accounting profit in all three cases. Because the implicit costs differ, however, the economic profit is different in each of the scenarios. Despite the positive accounting profit of $10,000 in Scenario 1, the economic profit is $0 due to the implicit cost of $10,000. An economic profit of zero might appear to be an unpleasant situation for the firm, but in fact it is the opposite. The firm is earning enough revenue to cover its explicit costs and the foregone salary and interest of the owner. That is, the owner could not earn more in any other line of business. It is said that the firm enjoys a **normal profit** equal to $10,000 in this case.

In Scenario 2, the economic profit is -$5,000. That is, the revenue the firm earns is not sufficient to cover both the explicit and implicit costs. Even though the firm earns a profit on paper (i.e., a positive accounting profit), in a real sense, the firm is losing money. If the owner shut down the business and took her next best opportunity, she would earn an additional $5,000. As a result, the firm earns less than a normal profit in this case.

In Scenario 3, the economic profit is $5,000. The firm earns enough revenue to *more than* cover all explicit and implicit costs. In this case, if the owner shut down the business to produce elsewhere, she would actually lose $5,000. The firm clearly earns more than a normal profit in this industry. It should now be clear why neoclassical economists focus exclusively on economic profit. Economic profit is ultimately what affects a firm's decision to remain in an industry or to leave an industry. Since accounting profit is the same in all three scenarios, it is not a proper guide to managerial decision making.

Although it was not mentioned in Chapter 7, all the cost curves that we discussed in the last chapter include both explicit and implicit production costs. Unless otherwise noted, any reference to production cost in neoclassical theory should be understood to include both kinds of cost because those are the costs that influence firm behavior according to neoclassical economists.

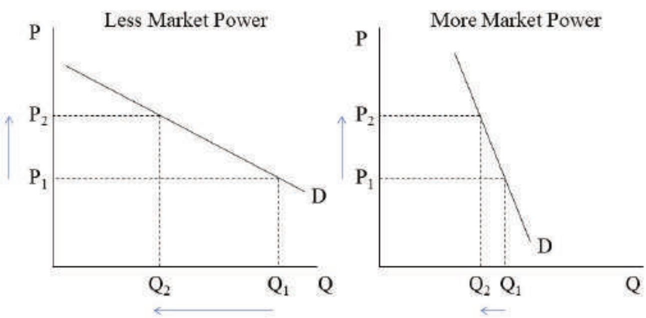
The Concept of Market Structure and the Meaning of Perfect Competition

We now wish to take a closer look at how the revenue measures defined in the last section behave as output changes. To accomplish this task, we must first discuss the concept of **market structure**. Market structure refers to all the characteristics of the marketplace that shape and influence how firms interact with their customers and with their competitors. Three key dimensions are used to distinguish between the different types of market structure:

1. The number and size of sellers
2. The ease of market entry and exit
3. The degree of product differentiation

How these characteristics combine in a specific market determines the degree of **market power** that each firm has in that market. Market power refers to the ability of a firm to raise the price of its product without losing all of its sales. A firm's market power is greater when the reduction in quantity demanded is smaller for a given increase in price. For example, Figure 8.1 shows two demand curves facing two different firms in two different markets for a similar good.

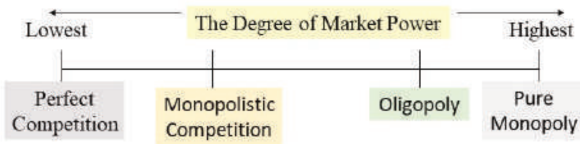
Figure 8.1: Differing Degrees of Market Power



The firm facing a flatter demand curve possesses less market power because the rise in price from P_1 to P_2 leads to a much sharper reduction in quantity demanded. The firm facing a steeper demand curve possesses considerable market power.

We can now refer to a **market power spectrum** that allows us to distinguish between four different types of market structure in terms of how much or how little market power the firms in each market structure have. Figure 8.2 represents the market power spectrum.

Figure 8.2: The Market Power Spectrum



Each of these market structures is distinguished along the lines of the three dimensions mentioned previously. The only market structure that we will examine in this chapter is **perfect competition**. According to neoclassical economists, a perfectly competitive market structure has the following three characteristics:

1. A large number of sellers and buyers
2. No barriers restrict the freedom of buyers or sellers to enter or exit the market
3. Each firm produces a homogeneous or standardized product

We should consider a few examples of actual markets that closely resemble perfectly competitive markets. For example, agricultural markets are highly competitive. The markets for wheat or corn have many buyers and sellers, and these crops are found to be very similar when we compare the product of one seller with that of another. It is also relatively easy for buyers and sellers to enter and exit these markets. Other examples include the markets for precious metals (e.g., gold and silver) and markets for corporate stock. In these markets, the standardization of the thing being sold is plainly seen.

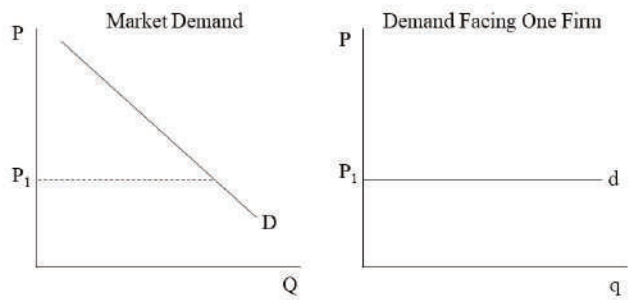
The implication is that no seller or buyer in a perfectly competitive market has any market power. That is, each seller (or buyer) is a **price-taker**, and so is powerless to change the market price. In all these markets, competition is so intense that no single buyer or seller has the power to raise or lower her price above or below the market price without reducing her sales to nothing (in the case of a price increase) or needlessly sacrificing revenue (in the case of a price decrease). The graphical analysis of a price-taking firm is taken up in the next section.

Neoclassical economists assert that the phrase "perfect competition" is entirely descriptive in nature. Indeed, these three characteristics appear to simply describe certain features of specific markets. It should be noted, however, that the perfectly competitive market structure is the *normative* standard in neoclassical economics as well. That is, it is the *moral* ideal toward which market capitalist economies should strive, according to this school of thought. The reason neoclassical economists are such strong advocates of perfect competition is that such markets can be shown to lead to *economic efficiency*, as defined in chapter 2. Later in this chapter, we will see how neoclassical economists arrive at this result.

The Revenue Structure of a Perfectly Competitive Firm

In order to determine a perfectly competitive firm's revenue pattern, we must first analyze the demand curve facing such a firm. From Chapter 3, we know that the market demand curve is downward sloping due to the law of demand, but the demand curve facing the individual firm is horizontal, as shown in Figure 8.3.

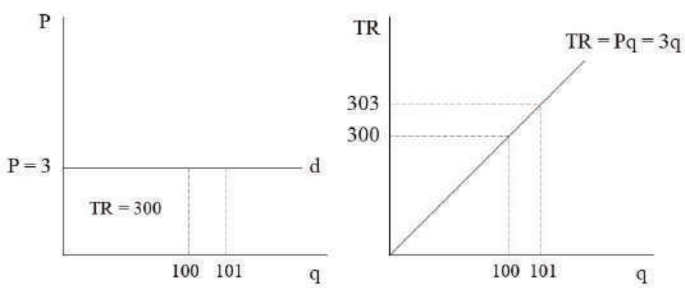
Figure 8.3: Market Demand and the Demand Facing a Perfectly Competitive Firm



The reason for the difference is that the individual perfectly competitive firm is a price-taker and only faces a tiny segment (d) of the entire market demand (D). If the perfectly competitive firm attempts to raise the price above P_1, then the quantity demanded (q) will fall to 0 units because every buyer can obtain a perfect substitute from a competitor at a price of P_1. Additionally, it makes no sense for the firm to reduce the price below P_1 because the firm can sell as much as it wants at the market price. It will sacrifice revenue needlessly if it cuts price because a price cut is not necessary to sell additional units. The firm is so small relative to the entire market that plenty of customers exist at the market price. We can also state that demand is infinitely elastic in the case of the demand curve facing the perfectly competitive firm. The smallest price increase will cause quantity demanded to fall to zero, and the smallest price cut will cause quantity demanded to soar.

We can now introduce the total revenue (TR) curve for a perfectly competitive firm. Figure 8.4 shows the TR curve for a perfectly competitive firm that faces a constant market price of $3 per unit.

Figure 8.4: The Total Revenue Curve



The graph on the left in Figure 8.4 shows that the area of the box under the demand curve is equal to total revenue since it is calculated as the product of price and quantity demanded. Furthermore, an additional unit can be sold at a price of $3 because the firm is a price-taker. The graph on the right shows what happens to total revenue as the quantity sold rises. It increases in a linear fashion because with each additional unit sold, the TR rises by the amount of the price of that unit, which is constant since the firm is a price-taker. Hence, the 101^{st} unit raises TR from \$300 to \$303.

We can also see that the TR curve is linear because TR = Pq, and the price is constant. Hence, the TR curve will have a zero intercept and a constant slope. If the price rises, then the TR curve will still rise from the origin, but the line will become steeper. That is, TR will rise more quickly as quantity rises. On the other hand, if the price decreases, then the TR curve will become flatter, indicating that TR rises more slowly. In addition, a price increase will shift the demand curve facing the firm upwards, and a price reduction will shift the demand curve facing the firm downwards.

Next, we wish to investigate the behavior of marginal revenue as the quantity sold changes. In the case of a price-taking firm, a unit increase in quantity always increases TR by the amount of the price. Hence, the MR is always equal to the price in the case of a perfectly competitive firm. That is:

$$MR = \frac{\Delta TR}{\Delta q} = P$$

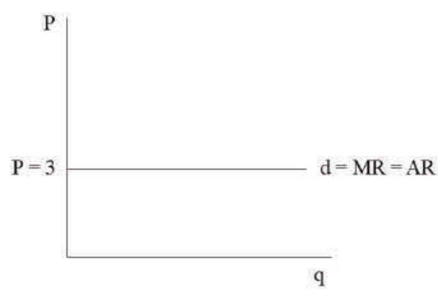
In Figure 8.4, for example, the MR of the 101^{st} unit is \$3 per unit. Also, the slope of the TR curve is equal to the constant price, which in turn is equal to ΔTR/Δq. Hence, mathematically, we can see that MR = P. The MR curve is, therefore, the same as the horizontal demand curve facing the firm.

We can also consider the behavior of average revenue as the price changes. The AR in the case of a perfectly competitive firm is equal to the price as shown below:

$$AR = \frac{TR}{q} = \frac{Pq}{q} = P$$

In Figure 8.4, for example, the AR when 100 units are sold is equal to $300/100 units or $3 per unit. Similarly, the AR when 101 units are sold is equal to $303/101 units or $3 per unit. Since AR = P, the AR curve is also the same as the horizontal demand curve facing the perfectly competitive firm. Figure 8.5 shows the MR and AR curves on a graph.

Figure 8.5: The Marginal Revenue and Average Revenue Curves



All the revenue measures can also be expressed in tabular form as shown in Table 8.2.

Table 8.2: Revenue Measures for a Perfectly Competitive Firm

Price (P)	Quantity (q)	Total Revenue (TR)	Marginal Revenue (MR)	Average Revenue (AR)
$3	0	$0	--	--
3	1	3	$3	$3
3	2	6	3	3
3	3	9	3	3
3	4	12	3	3
3	5	15	3	3
3	6	18	3	3

Methods of Profit Maximization in the Short Run

The behavioral assumption that neoclassical economists impose is that all firms seek to maximize economic profit. A considerable amount of disagreement exists as to whether profit maximization is the primary objective of firms. Some critics argue that managers of modern corporations pursue revenue growth and market share much more aggressively than maximum profits. Others argue that firms consider their broader sense of social responsibility, which includes a commitment to the firm's stakeholders (e.g., customers, the community, employees) rather than simply a commitment to the firm's shareholders. Finally, it might be argued that firms balance multiple and competing objectives in their operations and that no single goal should be elevated above the others. Whatever may be the case, we will assume that the firm strives only to maximize its economic profit, in accordance with the neoclassical theory of the firm.

Because the perfectly competitive firm takes the market price as given, it need not ask which price is the best price to charge. In the short run, the only decision the firm must make is how much output to produce to maximize its economic profit. One way for the firm to achieve this goal is to compare TR and TC at each output level. Because the difference is the total economic profit (π), the firm can simply select the output level that maximizes that difference. Table 8.3 provides an example of a firm that aims to solve this profit maximization problem in the short run.

Table 8.3: Maximization of Total Economic Profit in the Short Run

P	q	TR	TVC	TFC	TC	π
3	0	0	0	300	300	-300
3	100	300	40	300	340	-40
3	200	600	70	300	370	230
3	300	900	90	300	390	510
3	400	1200	140	300	440	760
3	500	1500	240	300	540	960
3	**600**	**1800**	**530**	**300**	**830**	**970**
3	700	2100	900	300	1200	900
3	800	2400	1700	300	2000	400
3	900	2700	2700	300	3000	-300
3	1000	3000	3800	300	4100	-1100

According to Table 8.3, the firm will maximize its economic profit by producing 600 units of output. At that output level, its economic profit will be $970. One can also see that if the firm produces too little or too much, then it will experience economic losses. Producing too little means that the firm fails to take sufficient advantage of labor specialization and the division of labor. Producing too much means that the firm fails to recognize the negative impact that diminishing returns to labor carries for profitability.

Figure 8.6 shows two ways of graphically representing the maximum economic profit.

Figure 8.6: Graphical Representations of Maximum Economic Profit

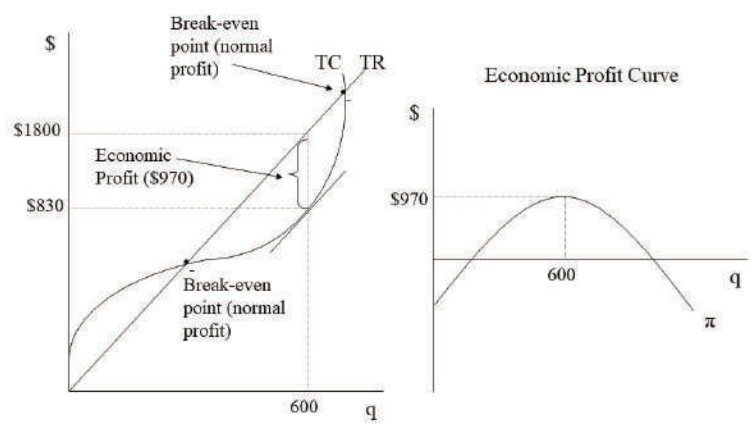


Figure 8.6 shows that economic losses exist to the left of the first break-even point because TC exceeds TR. At the break-even point, TC = TR and so economic profit is $0. Earlier in the chapter, it was argued that an economic profit of zero is still acceptable to the firm because all costs are covered. Because the opportunity costs are covered as well, the firm earns a normal profit. In between the two break-even points, the firm's revenues exceed its costs. Therefore, positive economic profits are

earned over that range of output. At a single output level, however, the gap between the TR and the TC is maximized. At that point where Q = 600 units, the economic profit is at a maximum.

The graph on the right represents the economic profit curve. It measures the difference between the TR and the TC, as shown in the graph on the left. The break-even points occur where the economic profit curve crosses the horizontal axis. Clearly, it reaches its maximum where Q = 600 units.

A second way to determine the profit-maximizing approach is less obvious but more useful. Managers generally do not have access to information about TR and TC at every possible output level. Instead, they base their production decisions on how profits rise or fall with small adjustments to output. This information is contained in the MC and MR measures, which have been calculated and are shown in Table 8.4.

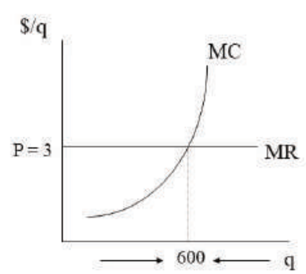
Table 8.4: The Marginalist Approach to Profit Maximization

q	MR	MC
0	--	--
100	3	0.40
200	3	0.30
300	3	0.20
400	3	0.50
500	3	1.00
600	3	2.90
700	3	3.70
800	3	8.00
900	3	10.00
1000	3	11.00

As Table 8.4 shows, MR is constant and equal to the product price of $3 per unit. The MC, on the other hand, falls and then rises for reasons already explained in chapter 7. We might start by asking whether the firm would find it profitable to produce the 100th unit of output. Since the sale of that additional unit will generate $3 of additional revenue and its production will add only $0.40 to cost, that unit will clearly add to the firm's economic profit. The 200th unit will also add to the firm's economic profit because the MR of $3 per unit clearly exceeds the MC of $0.30 per unit. In fact, the firm will continue to increase its output as long as MR exceeds MC. Once the firm reaches the 600th unit, the MR of $3 just barely exceeds the MC of $2.90 per unit and so it will be produced. If the firm increases its output any further, however, the MC of $3.70 will exceed the MR of $3 per unit and so it will not be profitable to produce that unit. The firm should, therefore, stop production at 600 units.

Even though MR and MC are not exactly equal at 600 units, they are close to equal. If we were to show more data, we could imagine increasing the output level somewhat above 600 units (but not as high as 700 units) until we reach the point where MR exactly equals MC. We know that MC will continue to rise due to diminishing returns to labor. The law of diminishing returns is the reason that the firm ceases production at a specific level of output in the short run. Beyond a certain point, the upward pressure on unit cost simply becomes too great. Figure 8.7 provides a graphical representation of how marginal adjustments to output can be used to determine the profit maximizing level of output for a perfectly competitive firm in the short run.

Figure 8.7: Profit Maximization Using Marginal Adjustments to Output



We can summarize how these marginal adjustments lead to the profit-maximizing choice of output:

1. If MR > MC, then the firm should increase its output.
2. If MR < MC, then the firm should reduce its output.
3. If MR = MC, then the firm should neither increase nor decrease its output.

It should also be noted that P = MR in the case of the perfectly competitive firm. Therefore, the rule that MR = MC can be written as P = MC for the perfectly competitive firm. Indeed, this condition is the first rule of profit maximization when we use the marginal approach.

Although the first rule of profit maximization is a necessary condition to ensure that profits are maximized in the short run, it is not a sufficient condition. That is, some instances arise in which a perfectly competitive firm would earn a greater profit by shutting down and producing zero units of output than by producing at the output level at which P = MC.

When would the firm decide to shut down? The second rule of profit maximization states that a firm will shut down when the product price falls below the firm's average variable cost (AVC). Another way of stating this rule is to state that the firm will only produce at the output level where MR = MC when the product price is at least as great as AVC. That is, the firm will only produce when P ≥ AVC. This rule might seem entirely arbitrary to the reader, but it can be proven with the help of a little basic algebra.

As we have seen, if the firm operates then its profits from operating (π_O) are the following:

$$\pi_o = TR - TC$$
$$\pi_o = TR - TVC - TFC$$

On the other hand, if the firm shuts down then its profits from shutting down (π_{SD}) are also:

$$\pi_{SD} = TR - TC$$
$$\pi_{SD} = TR - TVC - TFC$$

We can say more about the profits from shutting down. If the firm produces no output, then its revenues are zero. Furthermore, its variable costs are $0 because the firm will not purchase any labor. Hence, the profits from shutting down may be written as:

$$\pi_{SD} = Pq - wL - TFC = P(0) - w(0) - TFC$$
$$\pi_{SD} = -TFC$$

We can now compare the firm's profits from operating with the firm's profits from shutting down. In fact, the firm will only operate if the profits from operating are at least as great as the profits from shutting down. That is, the firm should operate if and only if $\pi_o \geq \pi_{SD}$. We can now derive the second rule of profit maximization as follows:

$$\pi_o \geq \pi_{SD}$$
$$TR - TVC - TFC \geq -TFC$$
$$TR - TVC \geq 0$$
$$TR \geq TVC$$
$$Pq \geq TVC$$
$$P \geq \frac{TVC}{q}$$
$$P \geq AVC$$

What this condition means is that the firm must earn enough revenue to cover its variable costs. If it does not earn this much revenue, then it makes more sense for the firm to shut down. Shutting down will cause the firm to lose its revenue, but firing all the workers will also allow the firm to eliminate its variable costs. The firm's loss will then be reduced to its TFC.

Five Possible Cases of Short Run Profit Maximization

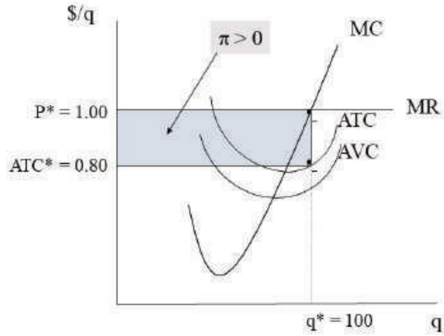
Given the marginal approach to short run profit maximization, we can identify five possible cases that might arise depending on the magnitude of the product price. Each of these cases as well as their implications for profitability are as follows:

1. $P > ATC \Rightarrow \pi > 0$
2. $P = ATC \Rightarrow \pi = 0$
3. $ATC > P > AVC \Rightarrow 0 > \pi > -TFC$
4. $P = AVC \Rightarrow \pi = -TFC < 0$
5. $P < AVC \Rightarrow \pi < -TFC < 0$

In each of the five cases, we will consider how the two rules of short run profit maximization influence the graphical analysis of the situation. Again, the firm should 1) produce where MR = MC conditional upon 2) P ≥ AVC. Otherwise, the firm should shut down and produce zero units of output.

Case 1 is the case of a positive economic profit as shown in Figure 8.8.

Figure 8.8: Case 1 – A Positive Economic Profit (P > ATC)

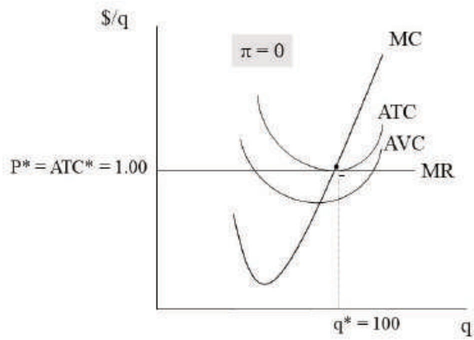


We begin our analysis of this case by finding the point at which MR = MC. Moving straight down from the intersection of these two curves to the horizontal axis gives us an output level of 100 units. Before we declare this output level to be the profit-maximizing choice, we need to check whether P ≥ AVC. In this case, it is greater. Even though the specific AVC has not been identified in the graph, we know that P exceeds AVC because at q* the MR curve is higher than the AVC curve. Therefore, the profit-maximizing output level is 100 units.

We can also use the information in Figure 8.8 to calculate TR, TC, and π. For example, TR is calculated as Pq so in this case TR is equal to $100 (= $1.00 per unit times 100 units). The TC is calculated as the product of ATC and q. To understand the reason, just recall that ATC = TC/q. Therefore, TC may be written as ATC·q. In this case, TC is equal to $80 (= $0.80 per unit times 100 units). We can now determine the total economic profit as the difference between TR and TC. In this case, π = \$20 (= \$100 – \$80) and is simply the area of the shaded region in the graph.

Case 2 is the case of an economic profit of zero as shown in Figure 8.9.

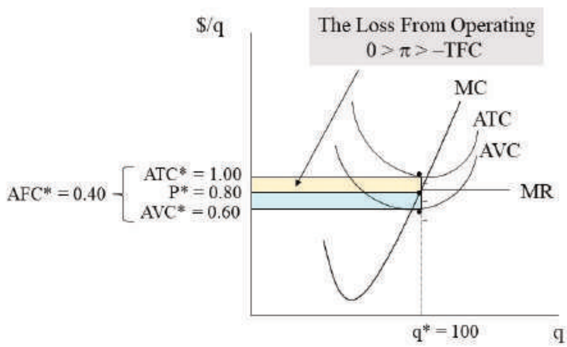
Figure 8.9: Case 2 – The Break-Even Point (P = ATC)



Again, we find the intersection of MR and MC to occur at 100 units of output. In addition, at that output level, the MR curve lies above the AVC curve and so the firm will operate. The profit-maximizing output level is 100 units as before. The TR in this case is \$100 (= \$1.00 per unit times 100 units). Because the ATC is also \$1.00 per unit, the TC in this case is \$100 as well. As a result, the economic profit is zero. Another way to see why economic profit is equal to zero is to notice that the box representing TR in the graph also represents TC. The reader should recall that the firm earns a normal profit in this case, and so this situation is not unacceptable to the firm. That is, the firm could not earn a greater profit in any other industry.

Case 3 is the case of a firm that suffers an economic loss but chooses to operate anyway as shown in Figure 8.10.

Figure 8.10: Case 3 – Operating at a Loss (ATC > P > AVC)

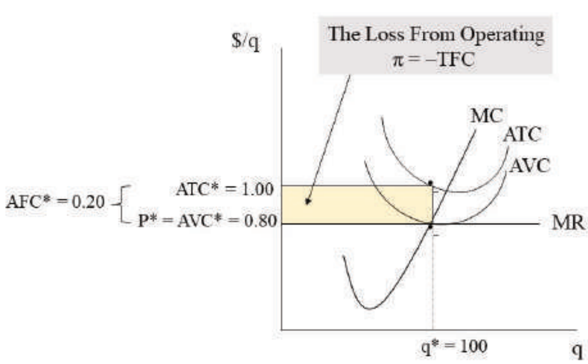


That is, shutting down is an even worse option. Let's see why. The intersection of MR and MC occurs at 100 units of output as before and P = $0.80 is clearly above AVC = $0.60 at that output level. Hence, the firm will operate. The firm's TR is equal to $80 (= $0.80 per unit times 100 units) and the firm's TC is equal to $100 (= $1.00 per unit times 100 units). The firm's economic profit is, therefore, equal to -$20 (= $80 minus $100). This loss is represented in Figure 8.10 as the top shaded box.

Why would the firm operate in this situation in the short run? If the firm shuts down, we know that its economic profit is always equal to – TFC. We can show the TFC graphically as the product of AFC and q. The reason is that AFC = TFC/q. Therefore, TFC = AFC·q. Furthermore, since ATC = AFC+AVC, the AFC = ATC – AVC. Hence, we can calculate the AFC in this case to be $0.40 per unit (= $1.00 per unit – $0.60 per unit), and the TFC is then equal to $40 (= $0.40 per unit times 100 units). Therefore, the economic profit from shutting down (= – TFC) is -$40. Clearly, this loss is much greater than the loss from operating. The loss from shutting down is equal to the two shaded areas combined. It follows that the firm will operate to maximize its economic profit. Certainly, the two rules of profit maximization led us to this conclusion much more quickly!

Case 4 is the case of a firm that is indifferent between operating and shutting down because its profit/loss situation is the same in either case. Figure 8.11 represents this case.

Figure 8.11: Case 4 – The Shutdown Point (P = AVC)

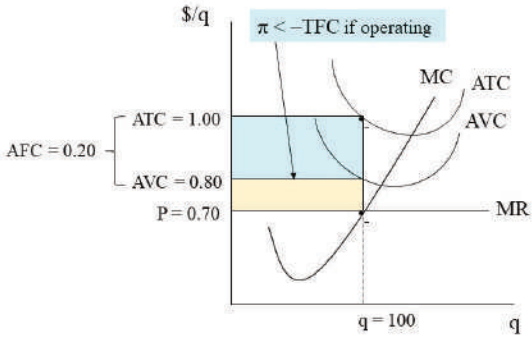


In this case, the intersection between MR and MC occurs at 100 units of output as before. The price, however, is exactly equal to AVC at this output level. The second rule of profit maximization requires that price be greater than *or equal to* AVC for the firm to operate. This condition is fulfilled and so the firm will operate. We can also see that TR is equal to $80 (= $0.80 per unit times 100 units) and TC is equal to $100 (= $1.00 per unit times 100 units). The firm's economic profit is, therefore, equal to -$20 (= $80 – $100). In addition, because the AFC = $0.20 per unit, the TFC = $20 (= $0.20 per unit times 100 units). If the firm shuts down then, its economic profit will be equal to -$20 (= –TFC). Clearly, the profit from operating is the same as the profit from shutting down. Because the firm is indifferent between operating and shutting down, by convention, we conclude that the firm will operate.

Students of economics are often puzzled by the conventional conclusion that a firm operates even though its economic profit is the same whether it operates or shuts down. In other words, why doesn't the owner just stay in bed if the profit/loss situation is the same either way? Although it may seem strange, the reader should remember that the revenue is sufficient to cover all costs, including the opportunity cost of operating the firm, which might include the value the owner places on additional sleep!

Case 5 is the only case in which the firm decides to shut down and produce zero units of output as shown in Figure 8.12.

Figure 8.12: Case 5 – The Shutdown Case (P < AVC)



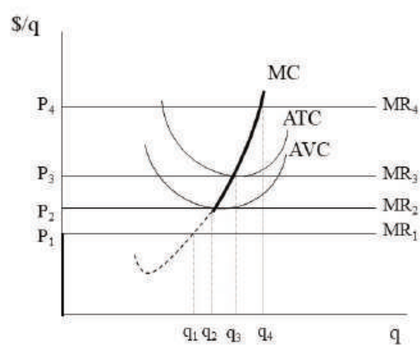
Again, the MR = MC intersection occurs at 100 units of output, but this time, the price of $0.70 per unit is below the AVC of $0.80 per unit. The second rule of profit maximization indicates that the firm should shut down in this case. The economic profit in this case is equal to –TFC. Because the AFC is $0.20 per unit, the TFC equals $20 (= $0.20 per unit times 100 units). The economic profit is, therefore, -$20. The careful reader might wonder how we can calculate TFC at an output level of 100 units when the firm has opted to produce zero units. The reason is that TFC is the same at all output levels so if we determine the TFC at 100 units of output, we also know the TFC at zero units of output.

It should also be noticed that if the firm had operated at 100 units of output, then its TR would equal $70 (= $0.70 per unit times 100 units) and its TC would equal $100 (= $1.00 per unit times 100 units). Its economic profit would then be -$30. That is, its economic loss from operating would exceed its economic loss from shutting down. In the graph, the economic loss from operating is equal to the sum of the two shaded regions. Only the top shaded region (equal to the TFC) is lost if the firm shuts down.

The Derivation of the Short Run Supply Curve

Now that we have shown how the perfectly competitive firm maximizes its economic profit in the short run, we can use the analysis to derive the firm's short run output supply curve. Figure 8.13 shows a series of MR curves corresponding to different prices as determined in the competitive market for this product.

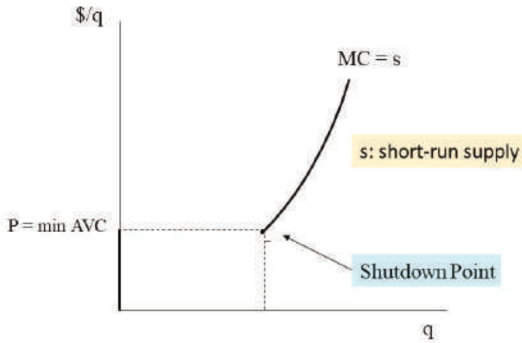
Figure 8.13: Marginal Cost and Short Run Output Supply



As the market price falls, the MR curve shifts downward. At the highest price of P4, the two rules of profit maximization indicate that q4 should be produced. At that output level, MR = MC and P > AVC. When the price declines to P_3, the firm will produce q_3, and the break-even point will be reached since P = ATC. When the price declines further to P_2, q_2 is produced and the shutdown point is reached since P = AVC. Finally, when the price falls to P_1, the firm maximizes its economic profit by shutting down and producing zero units of output because P < AVC. This analysis has allowed us to observe the different quantities of output supplied at different market prices, other factors held constant. In other words, the short run analysis of profit maximization has made possible the derivation of the firm's output supply curve. Because the intersections of MR with MC determine the profit-maximizing output levels, we can conclude that the supply curve is the MC curve above the minimum AVC. In Figure 8.13, the supply curve is the darkened portion of the MC curve plus the vertical axis below the minimum AVC because the firm produces zero units of output when the market price falls below that level.

Figure 8.14 shows the perfectly competitive firm's short run supply curve without the interference of the unit cost and MR curves.

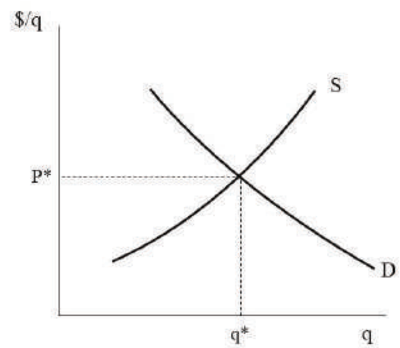
Figure 8.14: The Short-Run Supply Curve
of a Perfectly Competitive Firm



This analysis reveals the reason why we drew the supply curves in chapter 3 as suspended without a vertical intercept. We can also reinterpret the market supply curve. In chapter 3, it was explained that the market supply curve is the aggregation of many individual sellers' supply curves, which we can obtain through horizontal summation. We now see that the market supply curve is also the horizontal summation of the individual firms' MC curves.

At this stage, we have provided thorough explanations of the market supply and market demand curves. We have shown that production technology and capital endowments are the most important determinants of product market supply. The profit-maximizing behavior of perfectly competitive firms led us to this conclusion. Similarly, we have shown that consumer preferences and income endowments are the most important determinants of product market demand. The utility-maximizing behavior of individual consumers led us to this conclusion. As Wolff and Resnick explain, these fundamental determinants are used to provide the complete explanation for all commodity prices in capitalist economies within the neoclassical framework.[1] We now have a much deeper understanding of the graph in Figure 8.15 than we previously had.

Figure 8.15: Market Demand and Market Supply

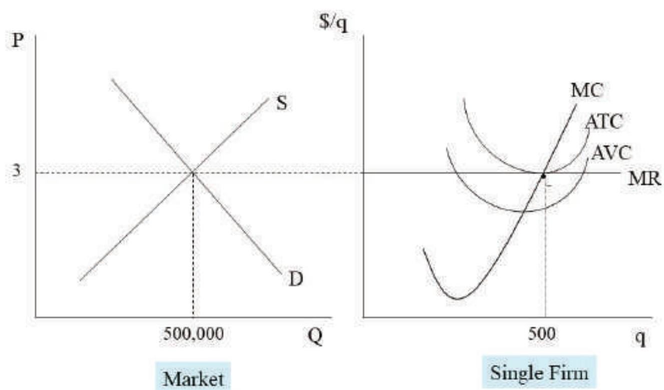


Long Run Profit Maximization

We now turn to an analysis of profit maximization in the long run when all inputs are variable. Because capital inputs are not fixed in the long run, firms will enter the industry if economic profits exist, and they will exit the industry if economic losses exist. For the purposes of this analysis, we will assume that all firms possess identical short run ATC curves. This assumption is reasonable so long as all firms have access to the same production technologies and face the same opportunity costs. At this stage, we will ignore the adjustments that firms make to their plant sizes and focus exclusively on the impact that the entry and exit of competing firms has on the profit/loss situations of firms in the industry.

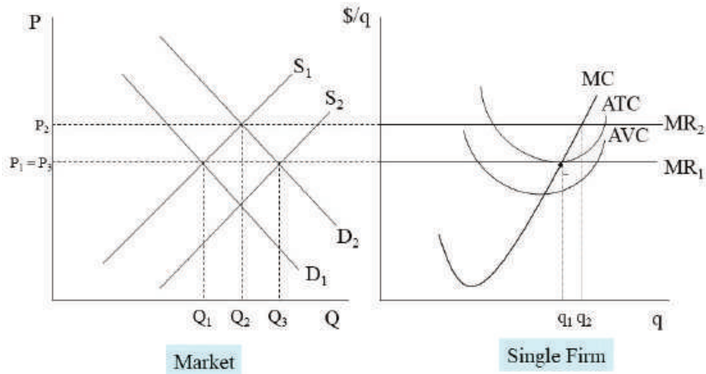
It turns out that the long run equilibrium outcome for a firm in a perfectly competitive market is the break-even case we considered in our analysis of the short run. That is, the market price is determined competitively through the interaction of supply and demand, and each firm earns an economic profit of zero as shown in Figure 8.16.

Figure 8.16: Long-Run Equilibrium



Why does this case represent the long run equilibrium outcome? The reason is that short run deviations from this situation produce an inherent long run tendency to change in the direction of this outcome. For example, in Figure 8.17, the market price and quantity exchanged begin at P_1 and Q_1, and the firm produces q_1 units of output.

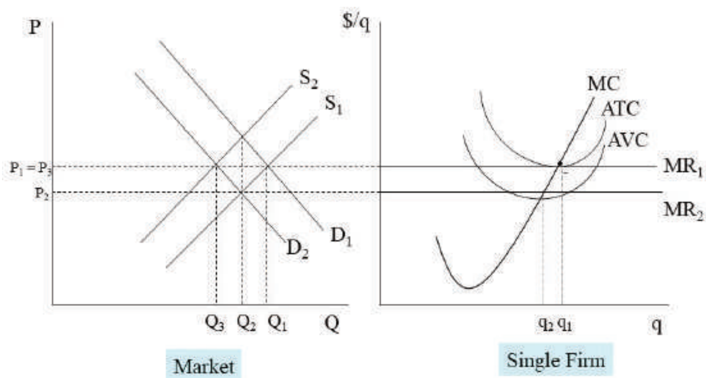
Figure 8.17: Adjustment to Long-Run Equilibrium – The Case of Profits



Now assume that a short run increase in demand causes the demand curve to shift to D_2. The increase in the market price to P_2 causes the firm's MR curve to shift upwards from MR_1 to MR_2. The firm's profit-maximizing choice of output, therefore, rises from q_1 to q_2. The price charged now exceeds ATC resulting in positive economic profits. In the long run, firms outside the industry respond to the resulting economic profits, which function as a signal that they should enter this industry. As the number of sellers in the industry increases, the market supply curve shifts rightward from S_1 to S_2. The market price falls to P_3, and the economic profits return to zero as the firm reduces output to the original level of q_1. It should be noted that the market price returns to its original level, but the market quantity exchanged has increased from Q_1 to Q_3. This result is argued to be consistent with Adam Smith's Invisible Hand in that the free market has reallocated a part of society's resources toward the production of this product for which consumer demand has grown.

Next assume that a short run decrease in demand causes the demand curve to shift to D_2 as shown in Figure 8.18.

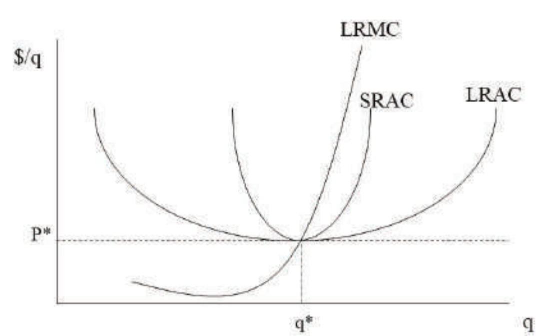
Figure 8.18: Adjustment to Long-Run Equilibrium – The Case of Losses



The decrease in the market price to P_2 causes the firm's MR curve to shift downwards from MR_1 to MR_2. The firm's profit-maximizing choice of output, therefore, falls from q_1 to q_2. The price charged is now below ATC resulting in economic losses. In the long run, firms inside the industry respond to the economic losses, which function as a signal that they should exit this industry. As the number of sellers in the industry decreases, the market supply curve shifts leftward from S_1 to S_2. The market price rises to P_3, and the economic profits return to zero as the firm increases output to the original level of q_1. It should be noted that the market price returns to its original level, but the market quantity exchanged has decreased from Q_1 to Q_3. This result is also argued to be consistent with Adam Smith's Invisible Hand in that the free market has reallocated a part of society's resources away from the production of this product for which consumer demand has fallen.

In the long run, firms also adjust their plant sizes, which is a feature we ignored when explaining the adjustment to long run equilibrium above. In chapter 7, we discussed the U-shaped LRAC curve, but we did not include the **long run marginal cost** (LRMC) curve. The LRMC curve intersects the LRAC curve at its minimum point as shown in Figure 8.19.

Figure 8.19: The Optimal Plant Size



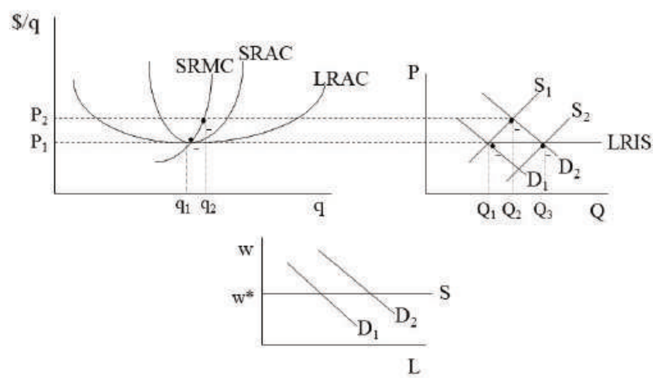
This result is expected because, as we have seen, whenever a marginal contribution is below an average, the average falls and whenever the marginal contribution is above an average, the average rises.

The firm will maximize long run economic profits by equating P and LRMC. Now suppose that the profit-maximizing choice is such that price exceeds LRATC. Then economic profits are made and firms will enter, driving the price down. As the price falls, the firm will reduce its plant size. Alternatively, suppose the profit-maximizing choice is such that the price is below LRATC. Then economic losses exist, and firms will exit the industry pushing the price up. As the market price rises, the firm will increase its plant size. Eventually, the firm will produce at output level q* where the market price is equal to LRMC and LRATC. In that situation, the firm is breaking even in the long run. It should also be noted that the plant size that corresponds to this output level allows the firm to produce at minimum SRATC and minimum LRATC. That is, the firm exhausts the gains from specialization and the division of labor in this plant but does not increase production so much that diminishing returns to labor begin to drive up per unit cost. Furthermore, it exhausts the gains from economies of scale but does not enter the region of diseconomies of scale. It is the **optimal plant size** for this reason.

The Derivation of the Long Run Industry Supply Curve

We can use our analysis of long run profit maximization to derive the **long run industry supply** (LRIS) curve.[2] The LRIS curve may possess different shapes depending on the way in which input prices respond as additional firms enter the market. If an abundance of the necessary inputs exists, then input prices might remain constant as additional firms enter the market. This type of industry is referred to as a **constant-cost industry**. Figure 8.20 shows how long run profit maximization may be used to derive the LRIS curve in a constant-cost industry.

Figure 8.20: The Long-Run Industry Supply Curve for a Constant-Cost Industry

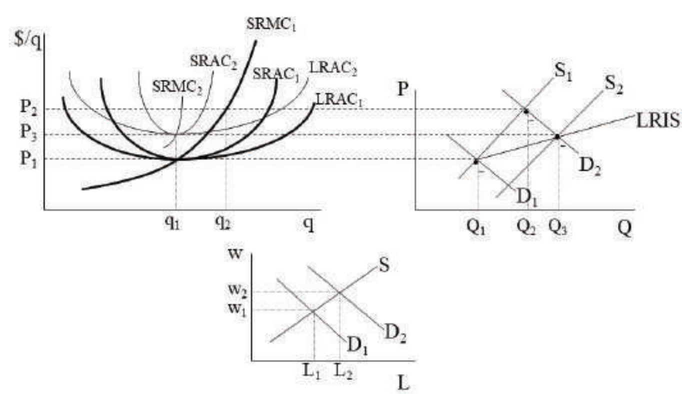


In Figure 8.20, a short run increase in demand from D1 to D2 causes an increase in the product price from P1 to P2. As a result, the firm expands its output in the short run from q1 to q2. The resulting economic profits cause competitors to enter the industry. As additional firms enter, they increase the demand for inputs, such as labor, but because input supplies are perfectly elastic (i.e., horizontal), input prices (e.g., wages) do not rise. As a result, the LRAC remains fixed. Market supply, therefore, increases from S1 to S2 until the price returns to its original level, and economic profits are again zero. Finally, we can connect the original equilibrium P1 and

Q_1 with the new equilibrium at P_1 and Q_3 with a straight line. This horizontal line is the LRIS curve for a constant-cost industry. It shows that the market can expand in the long run without any upward pressure on the product price, precisely because no upward pressure on the inputs prices occurs with the expansion.

If the necessary inputs are relatively scarce, then input prices are more likely to rise as additional firms enter the market. This type of industry is referred to as an **increasing-cost industry**. Figure 8.21 shows how long run profit maximization may be used to derive the LRIS curve in an increasing-cost industry.

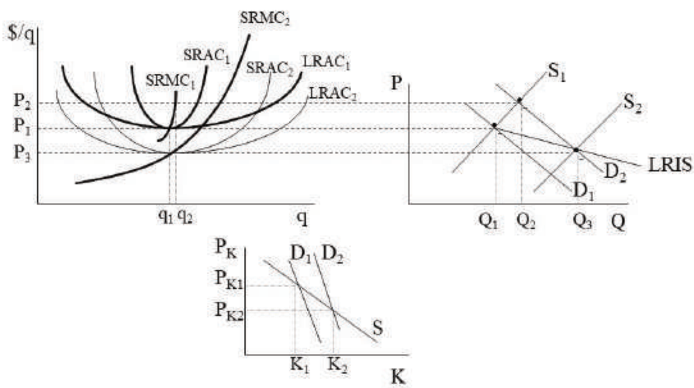
Figure 8.21: The Long-Run Industry Supply Curve for an Increasing-Cost Industry



In Figure 8.21, a short run increase in demand from D_1 to D_2 causes an increase in the product price from P_1 to P_2. As a result, the firm expands its output in the short run from q_1 to q_2. The resulting economic profits cause competitors to enter the industry. As additional firms enter, they increase the demand for inputs, such as labor, and because input supply curves are upward sloping, input prices (e.g., wages) rise. As a result, the LRAC shifts upward due to the rise in unit costs. Market supply, therefore, increases from S_1 to S_2 until the price declines to P_3, at which point economic profits are again zero. Finally, we can connect the original equilibrium P_1 and Q_1 with the new equilibrium at P_3 and Q_3 with a straight line. This upward sloping line is the LRIS curve for an increasing-cost industry. It shows that the market can expand in the long run only by putting upward pressure on the product price because upward pressure on the input prices occurs during the expansion.

If significant economies of scale exist in the input markets, then input prices may decline as additional firms enter the market. This type of industry is referred to as a **decreasing-cost industry**. Figure 8.22 shows how long run profit maximization may be used to derive the LRIS curve in a decreasing-cost industry.

Figure 8.22: The Long-Run Industry Supply Curve
for a Decreasing-Cost Industry



In Figure 8.22, a short run increase in demand from D_1 to D_2 causes an increase in the product price from P_1 to P_2. As a result, the firm expands its output in the short run from q_1 to q_2. The resulting economic profits cause competitors to enter the industry. As additional firms enter, they increase the demand for inputs, such as capital, but because input supply curves are downward sloping (reflecting scale economies), input prices (e.g., prices of capital goods) fall. As a result, the LRAC shifts downward due to the reduction in unit costs. Market supply, therefore, increases from S_1 to S_2 until the price declines to P_3, at which point economic profits are again zero. Finally, we can connect the original equilibrium P_1 and Q_1 with the new equilibrium at P_3 and Q_3 with a straight line. This downward sloping line is the LRIS curve for a decreasing-cost industry. It shows that the market can expand in the long run even as the product price falls because downward pressure on the input prices occurs during the expansion. This situation has occurred, for example, in the market for desktop and laptop computers due to effective utilization of economies of scale in the production of computer components as the market for computers has grown.

Implications and Criticisms of the Neoclassical Model of Perfect Competition

The major implication of the neoclassical model of perfect competition is that this market structure achieves *economic efficiency*. In chapter 3, it was explained that competitive market equilibrium leads to the full employment of scarce resources and allocative efficiency. That is, all resources are fully employed and marginal benefit equals marginal cost for each good produced when all markets clear. In that discussion, it was explained that the demonstration of least cost production would be postponed until this chapter. We can now see that long run equilibrium in a perfectly competitive market will lead to least-cost production. The long run equilibrium outcome leads to a price that is just equal to minimum long run average total cost. Hence, productive efficiency is achieved. Because full employment, least-cost production, and allocative efficiency are all achieved in a perfectly competitive market economy, neoclassical economists conclude that this market structure achieves economic efficiency. It is, therefore, the normative standard in neoclassical economics.

Nevertheless, we need to reflect on several major criticisms of the neoclassical model of perfect competition and its efficiency conclusion. First, market demand reflects willingness and ability to pay for the good or service. An individual may desperately need a specific good (e.g., a medication) but at the same time, she cannot afford to purchase the good at the equilibrium price. Because she is at a point on the demand curve that is below the equilibrium price, she will remain in need even when the market clears. To call the outcome an efficient one suggests that it is the most desirable outcome, but it ignores the possibility that many in need will not be able to obtain the product. The reader should recall that neoclassical economists do not distinguish between wants and needs. This criticism of the model is, therefore, a criticism that only heterodox economists emphasize.

A second major criticism of the neoclassical model of perfect competition is that social welfare is defined entirely in terms of the efficient use of resources to satisfy consumer wants. The level of satisfaction derived from productive work is completely ignored. For example, workers may be efficiently employed, but what if they are working 14-hour days over hot stoves or in dangerous coal mines and hate their jobs? Their satisfaction from living in such a society is diminished regardless of how much satisfaction they may derive from consuming material goods. Many workers may prefer to reduce their working hours even if it means reducing their material consumption, but in capitalist economies that choice may not exist. For many workers, the choice is to work long hours or not work at all. Even if part-time work is available, it typically is offered at low pay and without key benefits such as health insurance.

A final criticism of the neoclassical model of perfect competition is that it ignores power relationships. Conflicts between social classes are not possible in the model, which only acknowledges individuals interacting voluntarily in the free market. The notion that the relationship between employer and employee is conflict-ridden is entirely absent. All resource owners compete on the same level. The owners of capital and the owners of labor confront one another as equals in the perfectly competitive marketplace. All take market prices as given and differences in relative power simply do not exist. As we will see in the next section, the Marxian analysis of intensely competitive markets disagrees sharply with these conclusions.

The Marxian Theory of Competitive Profit Rate Equalization

Our purpose in this section is to consider how Marxian economists think about intensely competitive market capitalist economies. As we saw in Chapter 4, Marxian economists place primary emphasis on class conflict in their analyses of capitalism. To understand how this feature of capitalism is emphasized in their analysis of capitalist competition, we must first introduce a new concept that Marxian economists call the **organic composition of capital (OCC)**. The organic composition of capital refers to the degree of constant capital intensity present in a capitalist production process. That is, a capitalist production process that uses relatively more constant capital than variable capital has a higher organic composition of capital than another capitalist production process that uses relatively less constant capital compared with variable capital. To make this definition more precise, we can define the OCC as follows:

$$OCC = \frac{c}{c+v}$$

According to this definition, the OCC tells us the fraction of the total capital advanced that consists of constant capital. If the total capital advanced is $1000 and the constant capital advanced is $700, then the OCC is 70%. In the extreme case that no constant capital is advanced (c = 0), the OCC equals 0. On the other extreme, if the entire capital consists of constant capital (v = 0), then the OCC equals 1 or 100%.

We are now able to consider a simple example of a capitalist economy with only five industries. Table 8.5 contains hypothetical data for the five industries.

Table 8.5: Hypothetical Data for Five Industries

	Industry 1	Industry 2	Industry 3	Industry 4	Industry 5	Entire Economy
Constant Capital ($)	80	70	50	30	20	250
Variable Capital ($)	20	30	50	70	80	250
Capital ($)	100	100	100	100	100	500
Organic Composition of Capital	80.00%	70.00%	50.00%	30.00%	20.00%	50.00%
Rate of Surplus Value	75.00%	75.00%	75.00%	75.00%	75.00%	75.00%
Surplus Value ($)	15.00	22.50	37.50	52.50	60.00	187.50
Rate of Profit	15.00%	22.50%	37.50%	52.50%	60.00%	37.50%
Value of Commodities	115.00	122.50	137.50	152.50	160.00	687.50

In this example, each of the firms advances $100 worth of capital, but the organic compositions of capital are very different. Industry 1 has a very high OCC of 80% whereas Industry 5 has the lowest OCC of only 20%. In neoclassical terminology, Industry 1 is very "capital-intensive" and Industry 5 is very "labor-intensive." The rates of surplus value are also assumed to be the same across the industries. Specifically, the rate of surplus value (s/v) is assumed to equal 75% in each industry.[3]

These assumptions of different OCCs and uniform rates of surplus value lead to differences in the rates of profit across the sectors. Why? Recall that labor-power is the only commodity that creates new value (and thus surplus value) in the Marxian framework. When an industry uses relatively more variable capital, it will necessarily produce more surplus value (assuming the same degree of exploitation or rate of surplus value) across industries. That is the reason that Industry 5 produces the most surplus value and has the highest rate of profit. Similarly, it is the reason that Industry 1 produces the least surplus value and has the lowest rate of profit. As the reader can see by looking at what happens to the OCC and the rates of profit from Industries 1 through 5, as the OCC falls, the rate of profit rises.

Because capitalists measure the profitability of their activities using the rate of profit, this situation is highly unstable in a competitive capitalist economy. Think about it. Why would a capitalist want to invest $100 in Industry 1 and earn a 15% return when she can invest the same $100 in Industry 5 and earn a 60% return. Capital's search for the highest rate of profit should cause it to exit industries that have low profit rates due to high OCCs and to enter industries that have high profit rates due to low OCCs.

In everyday life, we do not observe profit rates that are much higher in industries that use relatively more variable capital (e.g., garment-making) and profit rates that are much lower in industries that use relatively more constant capital (e.g., the automobile industry). The problem we face then is to explain how the rates of profit equalize across industries so that the same amount of capital generates approximately the same return regardless of where it is invested. From a static perspective, the solution to this problem requires that we calculate the **general rate of profit (r)** for this capitalist economy. The general rate of profit that applies to all capital once the profit rates have equalized may be calculated by dividing the aggregate surplus value (S) for the entire economy by the aggregate capital advanced (C+V) for the entire economy. We can calculate the general rate of profit in this example in the following way:

$$r = \frac{S}{C+V} = \frac{187.50}{500} = 37.50\%$$

Because all capital should earn the same rate of profit in a competitive capitalist economy, we can multiply the amount of capital invested in each industry by r to obtain the **average profit** that will be appropriated in each industry. Because the capital invested in each industry is the same ($100) and the general rate of profit is the same (r), the profit appropriated in each industry is the same in this example. Table 8.6 adds two additional rows of information to the information from Table 8.5 that applies specifically to this question of profit rate equalization.

Table 8.6: Profit and Production Price Calculations for Five Industries

	Industry 1	Industry 2	Industry 3	Industry 4	Industry 5	Entire Economy
Constant Capital ($)	80	70	50	30	20	250
Variable Capital ($)	20	30	50	70	80	250
Capital ($)	100	100	100	100	100	500
Organic Composition of Capital	80.00%	70.00%	50.00%	30.00%	20.00%	50.00%
Rate of Surplus Value	75.00%	75.00%	75.00%	75.00%	75.00%	75.00%
Surplus Value ($)	15.00	22.50	37.50	52.50	60.00	187.50
Rate of Profit	15.00%	22.50%	37.50%	52.50%	60.00%	37.50%
Value of Commodities	115.00	122.50	137.50	152.50	160.00	687.50
Profit ($)	37.50	37.50	37.50	37.50	37.50	187.50
Price of Production of Commodities ($)	137.50	137.50	137.50	137.50	137.50	687.50

Table 8.6 shows that the profit in each industry is $37.50. The reader should notice that the concept of profit is one that we have not yet discussed in the Marxian framework. It is not the same as surplus value. In fact, the equalization of the profit rate has created a situation in which the profits received and the surplus values produced are different in most of the industries in the table. In addition, the selling prices of the commodities will now be different from the values of the commodities. That is, the **prices of production** in each industry are equal to the total capital advanced in that industry plus the profits received in that industry.

From a dynamic perspective, we can say that capital flows out of low profit industries and into high profit industries. As capital moves, production in high profit industries increases. The increased production causes the market price of the commodities produced in those industries to fall below the market value of those commodities.[4] As a result, the profit rates fall in high profit industries. Similarly, production in low profit industries contracts as capital flows out of them. The decrease in production causes the market price of the commodities in those industries to rise above the market value of those commodities. As a result, the profit rates increase in low profit industries. Capital movements will continue to occur until the profit rates have equalized across all industries. In addition, we can see from Table 8.6 that prices of production rise above their values in low profit industries and fall below their values in high profit industries. Only Industry 3, which possesses the same OCC as the entire economy, earns a profit equal to the surplus value it produces and charges a price of production equal to the value of its commodities.

It is important not to become lost in the details and miss the fact that this analysis of competitive capitalism assigns a central role to class exploitation. It is the working class that produces the surplus value. Even though each industry ends up earning an amount of profit that is different from the surplus value it produces (except for Industry 3), the aggregate surplus value for the entire economy ($187.50) is equal to the aggregate profit. In other words, the capitalist class shares equally in the mass of surplus value produced by the working class according to each capitalist's share of the total capital. In addition, the aggregate value of the commodities produced depends upon the SNALT required for their production within the Marxian framework. As Table 8.6 shows, the aggregate value of $687.50 is the same as the aggregate production price of the commodities produced for the entire economy. Marx identified these two key aggregate equalities in volume III of *Capital* which we can summarize as follows:

$$Aggregate\ Surplus\ Value = Aggregate\ Profit$$

$$Aggregate\ Value = Aggregate\ Price\ of\ Production$$

The major objection to this Marxian analysis has been dubbed the **Transformation Problem**. The problem is easy enough to grasp but has proven incredibly difficult to solve to the satisfaction of all those interested in this question. The problem is that the elements of constant and variable capital also have values that must be *transformed* into production prices. In our example, these values have been left in their original form, giving rise to the objection that the transformation of values into production prices is not complete. For example, in Industry 1 constant capital might be advanced for the purchase of iron ore to be used in production. It will be purchased at its value, but once the profit rate has equalized, it should be purchased at its production price rather than its value. Many proposed solutions to this problem have been put forward since the late nineteenth century. The problem is that no one has yet discovered a way to transform the input values into production prices while at the same time maintaining both of Marx's aggregate equalities. This problem continues to be one of the most challenging problems in the history of economic thought. Despite this problem, Marxian analysis provides a powerful explanation for the equalization of profit rates in capitalist economies within the context of a framework that grants a central place to the class struggle between workers and capitalists.

Following the Economic News

On October 10, 2008, a metal manufacturer with production plants in Cranston, Rhode Island and Mishawaka, Indiana decided to shut down its operations.[5] The company, called Scott Brass, chose to shut down due to a sharp drop in the price of copper. The reduction in the price of copper was driven by the global economic downturn that began in 2007, which adversely affected housing construction. Because a large amount of copper electrical wire is used in the production of new homes, the drop in housing construction greatly reduced the demand for copper. The price of copper thus fell significantly. The closing of Scott Brass's plants led to a large number of layoffs. As many as 50 to 60 unionized workers, including drivers and manufacturing employees, were laid off due to the closing of the Cranston plant. In Indiana, the plant shutdown led to the loss of 78 jobs.

In terms of the analysis of perfect competition discussed in this chapter, the plant shutdowns can be understood to be the result of the price of copper products falling below the average variable cost of production in these production plants. As we have seen, once the market price falls below AVC, the firm can reduce its losses to the amount of the total fixed cost. By firing the workers, the firm loses the revenue from any metal products that would have been produced and sold, but the savings resulting from the firm no longer being required to pay wages more than makes up for the lost revenue. The firm would rather suffer the loss of its TFC in the short run than operate at a greater loss. The workers, of course, would prefer that the company continue to operate, but the owners are only interested in maximizing economic profit. We see here the potential for a conflict between the working class and the capitalist class over the control of the means of production, which is a characteristic of the Marxian analysis of competitive markets that we discussed. It should also be noted that the neoclassical analysis predicts the possible reopening of Scott Brass's production plants if the price of copper products rises above AVC before the firm is able to withdraw its capital investment.

Summary of Key Points

1. Economic profit is typically smaller than accounting profit because economic profit subtracts both explicit costs and implicit costs from total revenue.
2. A firm earns a normal profit when its economic profit is zero.
3. Neoclassical economists regard perfectly competitive markets as the normative ideal among market structures.
4. Firms in a perfectly competitive market structure are price-takers, and so average revenue and marginal revenue are equal to price.
5. To maximize economic profit in the short run, a perfectly competitive firm must produce such that price equals marginal cost and should only produce a positive output level when price is greater than or equal to average variable cost.
6. The short run supply curve of a perfectly competitive firm is the marginal cost curve above the minimum average variable cost of production.
7. In the long run, economic profits will be driven to zero as firms enter due to positive short run economic profits and as they exit due to negative short run economic profits.
8. The long run industry supply curve may be positively sloped, negatively sloped, or horizontal depending on the way in which input prices react as new firms enter an industry.
9. Neoclassical economists conclude that perfect

competition leads to economic efficiency, but they do not emphasize the distinction between market demand and needs, the power struggle between social classes, and the fact that people's happiness might depend as much on the work that they do as on the material consumption they enjoy.

10. In Marxian analysis, rates of profit differ across industries due to differing organic compositions of capital.
11. Capital moves between industries in search of the highest profit rate until production prices are formed, and the general rate of profit is established across all sectors.
12. The Transformation Problem exists because many believe that the values of inputs must be transformed into production prices in the same way that the values of final commodities are transformed into production prices.

List of Key Terms

Average revenue (AR)

Marginal revenue (MR)

Accounting profit

Explicit costs

Implicit costs

Economic profit

Normal profit

Market structure

Market power

Market power spectrum

Perfect competition

Price-taker

Long run marginal cost (LRMC)

Optimal plant size

Long run industry supply (LRIS)

Constant-cost industry

Increasing-cost industry

Decreasing-cost industry

Organic composition of capital (OCC)

General rate of profit (r)

Average profit

Prices of production

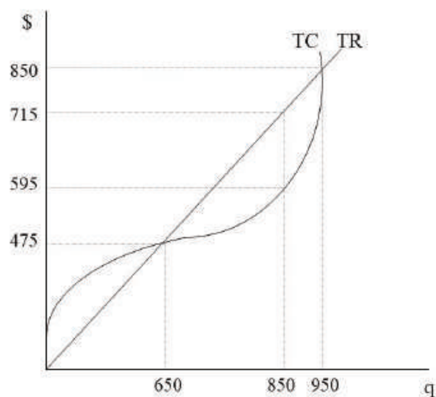
Transformation problem

Problems for Review

1.Suppose the market for frozen pizza is perfectly competitive and the current market price of a frozen pizza is $5.75. On a graph, draw the total revenue curve facing one firm in the market as well as the new total revenue curve when the market price rises to $6.25 per frozen pizza. Place dollars ($) on the vertical axis and quantity (q) on the horizontal axis.

2. What is the economic profit of the firm represented in the graph in Figure 8.23? Assume the firm is maximizing its economic profit.

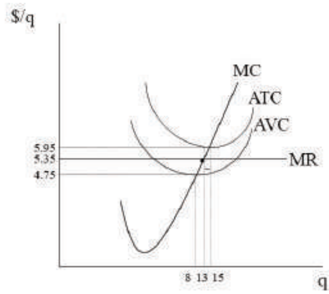
Figure 8.23: Problem 2



3. Answer the following questions when the firm represented in Figure 8.24 maximizes its economic profit.

- What is the profit-maximizing output level and price?
- What is total revenue?
- What is total economic cost?
- What is total economic profit?
- What is the total fixed cost?

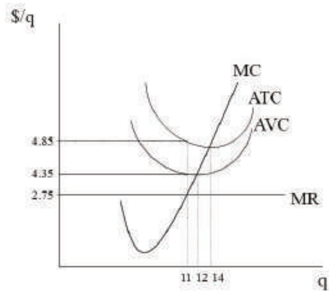
Figure 8.24: Problem 3



4. Answer the following questions when the firm represented in Figure 8.25 maximizes its economic profit.

- What is the profit-maximizing output level and price?
- What is total revenue?
- What is total economic cost?
- What is total economic profit?
- What is the total fixed cost?

Figure 8.25: Problem 4



5. Complete the remainder of Table 8.7 using the given information. What is the general rate of profit? Do Marx's two aggregate equalities hold?

Table 8.7: Problem #5

	Industry 1	Industry 2	Industry 3	Industry 4	Industry 5	Entire Economy
Constant Capital ($)	90	60	50	40	10	
Variable Capital ($)	10	40	50	60	90	
Capital ($)	100	100	100	100		
Organic Composition of Capital	90.00%	60.00%	50.00%		10.00%	
Rate of Surplus Value	50.00%	50.00%	50.00%	50.00%	50.00%	50.00%
Surplus Value ($)	5.00	20.00		30.00	45.00	
Rate of Profit	5.00%		25.00%	30.00%	45.00%	
Value of Commodities		120.00	125.00	130.00	145.00	
Profit ($)						
Price of Production of Commodities ($)						

Notes

1. Wolff and Resnick (2012), p. 38.
2. It was my experience as a teaching assistant for Prof. Thomas R. Swartz at the University of Notre Dame in the early 2000s that inspired my approach to the derivation of LRIS curves in this section.
3. In Table 8.5, the total surplus value may be calculated as the product of the rate of surplus value (s/v) and the variable capital (v). The rate of profit is calculated as s/(c+v) as in Chapter 4. Finally, the total value of the commodities produced in each industry is calculated as the sum of c, v, and s, which is also how the total value was calculated in Chapter 4.
4. Recall from Chapter 4 that fluctuations in supply and demand can cause the value of a commodity and its price to diverge.
5. Downing, Neil. "Scott Brass closes, putting more than 50 R.I. workers out of work." *The Providence Journal*. Projo.com. October 21, 2008. Web. Accessed on October 22, 2008.

CHAPTER 9

THEORIES OF MONOPOLISTIC MARKETS

Goals and Objectives:

In this chapter, we will do the following:

1. *Identify* the defining characteristics of monopolies and barriers to the entry of competitors
2. *Explain* the nature of the demand facing a monopolist and its marginal revenue
3. *Apply* the rules of short run profit maximization for a monopolist to a variety of situations
4. *Explore* the implications of monopoly for efficiency and the long run
5. *Analyze* the special cases of natural monopoly and price discriminating monopoly
6. *Investigate* the Austrian and Randian critiques of neoclassical monopoly theory
7. *Examine* the Marxian theory of monopoly capital

Neoclassical Monopoly Theory: Defining Characteristics and Types of Entry Barriers

In the nineteenth century, competition in many

American industries was quite fierce. As firms implemented new production technologies and new firms entered markets to capture a share of the profits, prices fell and the competitive struggle drove many firms out of business. In the iron and steel industry, for example, many iron and steel companies entered pricing pools in an effort to place limits on price competition. These agreements were not very stable, however, and by the late nineteenth century, a wave of mergers swept through American industry as firms strove for a way to stabilize their prices and profits. Out of this merger wave arose several giant corporations in a variety of different industries. These firms controlled such a large part of the total output of their industries that observers regarded them as **monopolies**. For example, the United States Steel Corporation (known simply as U.S. Steel) controlled most steel production in the United States after its formation in 1901. Other examples include International Harvester, which dominated the agricultural machinery market, and American Tobacco, which was dominant in the cigarette market. John D. Rockefeller established Standard Oil a few decades before the merger wave, and it was the dominant firm in the refining and transportation ends of the oil business.

Strictly speaking, a firm only qualifies as a monopoly firm if it is the *only* seller in a specific market. Technically, the firms that were born during the merger wave of the 1890s were not pure monopolies, but critics considered them monopolistic due to their dominance in their respective industries. The rise of large corporations encouraged neoclassical economists to develop an analysis of the behavior of the monopolistic firm. Before we examine this model of monopoly behavior in detail, we need to identify the defining characteristics of monopolies.

The first characteristic is that the seller must be the only seller in the entire market. On a related note, it is also important that no close substitutes exist for the product. If close substitutes do exist, then the market arguably has more than one seller because we should define the market broadly enough to include close substitutes. Finally, to ensure that the market remains dominated by a single seller, either natural or legal barriers must be in place to prevent potential competitors from entering the market. If all these conditions hold, then we regard the market as monopolistic.

It should now be clear that a monopoly market is the neoclassical antithesis of the perfectly competitive market. Without any competition from other sellers, the monopolist has the power to set the market price rather than taking it as given. It is a **price-setter** rather than a price-taker. Recall that market power is the power to increase the price of one's product without losing all of one's sales. The monopolist, therefore, has complete market power, limited only by consumer demand. That is, it can raise its price without the fear that competitors will charge lower prices, but it still must consider the fact that its customers will likely purchase less of its product at the higher price. In other words, the monopolist cannot subvert the law of demand.

Table 9.1 summarizes the differences between the perfectly competitive market and the monopolist market.

Table 9.1: The Characteristics of Perfectly Competitive Markets and Monopolistic Markets

Perfectly Competitive Markets	Monopolistic Markets
Many sellers	One seller
All firms sell a homogeneous product	No close substitutes are available
Easy entry and exit for sellers	Barriers to entry prevent sellers from competing
No firm possesses any market power	Maximum market power, limited only by demand

We need to explore the types of entry barriers that make monopolistic markets possible. Often the barriers are *legal* in nature. Legal barriers to entry include copyrights, patents, licenses, and the ownership of key inputs. We discuss each in turn.

A copyright refers to the exclusive right to reproduce materials (e.g., books, journals, magazines, musical recordings). Since 2003, copyright protection in the United States has been active during the life of the author plus 70 years. One widely cited example of copyright protection is Time Warner's ownership of the copyright to the song "Happy Birthday." As of 2004, Time Warner was earning about $2 million per year in royalties due to its ownership of this copyright.[1] It is the reason that the employees in many restaurants sing different songs rather than the song "Happy Birthday" to their customers when they learn that customers are visiting their restaurants on their birthdays.[2] Any public performance of the song without the payment of royalties to the owner of the copyright exposes the performer to a potential lawsuit. Another famous example is Michael Jackson's purchase of the copyright to the Beatles' songs in 1985 for $47.5 million.[3] Although it was a lot of money, the estimated value in 2006 was $1 billion![4]

Patents are another kind of legal barrier to entry. A patent is the exclusive right granted to an inventor to produce and sell a product for a specific period. In the United States, the government grants patent protection for 20-year terms. Patents are particularly important in the pharmaceutical industry in which large corporations develop certain types of drugs and medications and possess the exclusive right to sell them during the term of the patent. The justification for patents and copyrights is that firms and individuals need incentives to incur the costs of developing inventions and creating artistic and professional work. If competitors could immediately duplicate and compete in the sale of the newly invented product or the newly created work, then the creator would not have an opportunity to recoup the costs. The consequence would be a lack of invention and artistic activity in our society. The downside, of course, is that competition in the protected area does not allow the price of the new product or material to decrease. As a result, a tradeoff exists between innovation and low product prices. For this reason, patent and copyright protection have limited terms.

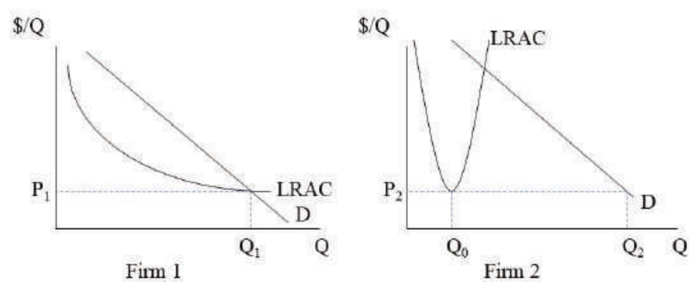
Licenses are another important type of barrier to entry. An occupational license is legal permission to conduct a specific line of business. For example, licenses are required to practice law and medicine. The belief is that some professions provide services that may harm consumers if unqualified individuals practice in those areas. The government uses licensing requirements then as a safeguard against the entry of unqualified service providers. Strictly speaking, we should not regard these markets as monopolistic. Those licensed to practice law *jointly* hold the monopoly power within the legal profession. Considerable competition exists between the many licensed professionals who sell legal services. The same applies to those who have licenses to practice medicine.

Finally, the ownership and control of key inputs is another legal barrier to entry because it depends on property rights. That is, if a firm has a monopoly in a key resource market, then it also frequently has a monopoly in the product market. For example, the Aluminum Company of America (ALCOA) once had a monopoly in the market for bauxite, which is the critical input to produce aluminum. Its monopoly in the market for bauxite allowed it to monopolize the aluminum market.[5] Similarly, in the late nineteenth century, Standard Oil acquired all the major pipelines for the transportation of oil in the United States helping it to solidify its monopolistic control of the transportation side of the oil business. During this same period, Carnegie Steel acquired a controlling interest in a major producer of coking coal, which the company then used in the production of steel. This acquisition helped Carnegie's company to dominate the steel market.

Aside from legal barriers to entry, we can also identify one key *economic* barrier to entry that may lead to the formation of a monopoly market. In Chapter 7, we discussed the concept of economies of scale (EOS). This concept refers to the way in which per unit production cost declines as a firm expands in the long run. This phenomenon occurs for a variety of reasons, including the spreading of large startup costs over many units of output, labor and managerial specialization, and learning by doing. Economies of scale may give one firm such a large economic advantage over its competitors that it is able to completely monopolize a market even in the face of unrestricted competition. When a firm acquires a monopoly due to economies of scale, neoclassical economists refer to it as a **natural monopoly**.

Consider the long run average cost (LRAC) curves of the two firms represented in Figure 9.1.

Figure 9.1: Economies of Scale as a Barrier to Entry



Firm 1 enjoys considerable economies of scale as can be seen from the falling LRAC over a large range of output. Firm 2, on the other hand, experiences diseconomies of scale at relatively low output levels as indicated by the rapid increase in LRAC even while the firm is operating at a very low output level. The market demand is similar in each market. If Firm 1 charges a low price (P_1) that just allows it to cover the minimum LRAC, then its production at the minimum efficient scale (MES) is sufficient to meet the entire quantity demanded in the market (Q_1) at that price. The situation is different with Firm 2 if it also sets a price (P_2) that just allows it to cover the minimum LRAC. Firm 2 will produce at MES, but this quantity (Q_0) is not nearly enough to supply the entire quantity demanded in the market (Q_2) at that low price.

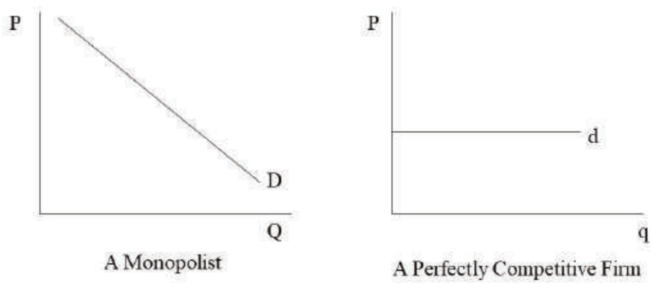
In this case, Firm 1 is the natural monopoly. The reason is that if multiple firms with the same LRAC curve were to produce only a fraction of the output, then the LRAC for each firm would be much higher than it is when a single firm produces for the entire market. A single firm in that case is the most efficient outcome from a cost-minimizing perspective, and the scale economies of the firm serve as a barrier to entry keeping out potential competition. On the other hand, if multiple firms produce in the industry in which Firm 2 operates, these firms can collectively produce enough to satisfy the entire quantity demanded in the market even as each produces at the MES. In that case, the most efficient outcome is to have multiple firms competing with one another, and this industry is not one in which a natural monopoly is likely to operate.

The Nature of Demand and Marginal Revenue in Neoclassical Monopoly Theory

We are now to the point where we can begin to analyze the neoclassical theory of monopoly markets. We begin by considering the demand curve facing the monopolist. Because the monopolist is the only seller in the market, the monopolist faces the entire market demand curve. Furthermore, this demand curve that the monopolist faces is likely to be relatively inelastic. The reason is that no close substitutes for this product exist, and we learned in Chapter 5 that the availability of substitutes is a major determinant of the price elasticity of demand. When no close substitutes are available, consumers become much less responsive to changes in the price of the product. This aspect of the market demand curve in a monopoly market strengthens the market power of the monopolist.

For the sake of comparison, we should consider how the demand curve facing the monopolist differs from the demand curve facing the perfectly competitive firm. Figure 9.2 shows the demand curves facing a monopolist and a perfectly competitive firm.

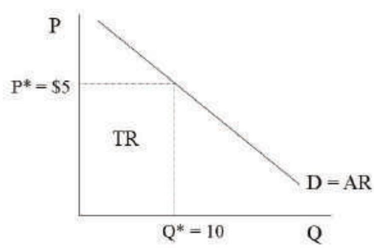
Figure 9.2: The Demand Curves Facing the Monopolist and the Perfectly Competitive Firm



The perfectly competitive firm faces a horizontal demand curve because the firm is small and only faces a segment of the market demand. As explained in Chapter 8, a perfectly competitive firm is a price-taker for this reason. Consumers are perfectly responsive to price changes and so demand is perfectly elastic.

As when we investigated the revenue structure of the perfectly competitive firm, we will now consider the revenue structure facing the monopolist. It is assumed at this stage that the monopolist charges the same price for every unit sold. As a result, the firm's total revenue (TR) may be represented in Figure 9.3 when the price charged is \$5 per unit and the quantity demanded is 10 units.

Figure 9.3: The Average Revenue Curve Facing a Monopolist



Clearly, the firm's TR in this case is \$50 (= \$5 per unit times 10 units) and is represented as the area of the rectangle in the graph. The average revenue (AR) can also be calculated as follows:

$$AR = \frac{TR}{Q} = \frac{\$50}{10\ units} = \$5\ per\ unit$$

It turns out that the AR in this case is the same as the price charged. This result will generally hold as can be seen below:

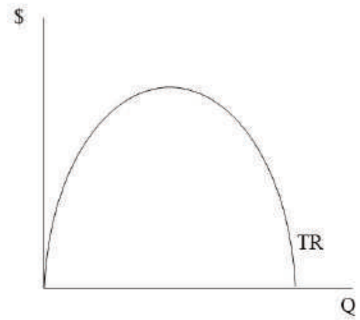
$$AR = \frac{TR}{Q} = \frac{PQ}{Q} = P$$

What this result implies is that at any level of output, we only need to look at the height of the demand curve to obtain the firm's AR just as we look at the height of the demand curve to obtain the price at that output level. Therefore, the demand curve facing the monopolist is identical to the firm's AR curve as indicated in Figure 9.3.

It should also be noted at this stage that the monopolist can set the product price or the quantity exchanged, but it cannot choose both variables independently. That is, if the monopolist selects the price, then the consumers will decide which quantity to demand at that price. Similarly, if the monopolist decides that it will sell a specific amount of its product, then the consumers will decide which price they are willing and able to pay to purchase that amount of the product.

We next turn to the monopolist's TR curve. As the reader might recall from Chapter 5, the TR curve has a specific shape in the case of a downward sloping linear demand curve. It is shown in Figure 9.4.

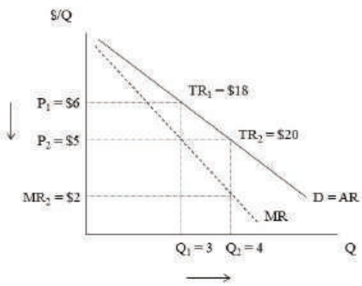
Figure 9.4: The Total Revenue Curve Facing a Monopolist



In Chapter 5, we learned that the reason for this shape is that at high prices demand is elastic and so price cuts lead to larger percentage increases in quantity demanded, which causes TR to rise overall. Once the price falls below the point of unit elasticity, demand becomes inelastic and further price cuts lead to smaller percentage increases in quantity demanded. As a result, TR falls. The pattern of TR as output changes is much more complicated for a monopolist than for a perfectly competitive firm. In the last chapter, we learned that a perfectly competitive firm's TR rises continuously in a linear fashion because the firm faces a constant price as output rises. In the case of the monopolist, the firm is cutting price to sell additional units of output, and the changing demand elasticity causes TR to rise and then fall.

The monopolist's marginal revenue (MR) curve is the final revenue concept that we must develop. Suppose that the monopolist faces the market demand curve shown in Figure 9.5 and that we know two points on the market demand curve.

Figure 9.5: The Marginal Revenue Curve Facing a Monopolist



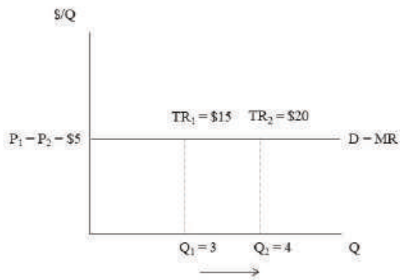
In Figure 9.5, the price falls from \$6 per unit to \$5 per unit, and the quantity demanded subsequently rises from 3 units to 4 units. The TR increases from \$18 to \$20 indicating that the firm is selling in the elastic portion of the market demand curve. The MR of the fourth unit of output may be calculated as follows:

$$MR = \frac{\Delta TR}{\Delta Q} = \frac{TR_2 - TR_1}{Q_2 - Q_1} = \frac{20-18}{4-3} = \$2 \; per \; unit$$

In other words, the firm's TR only rises by \$2 due to the increase in output from 3 to 4 units. This result might seem surprising, especially given the fact that the fourth unit of output sells at a price of \$5! What is the reason for this strange result? While it is true that the fourth unit sells at a price of \$5, it was necessary to reduce the price from \$6 to \$5 *for every unit sold* to sell that fourth unit. Therefore, the firm gains \$5 in additional revenue from the sale of the fourth unit even as it loses \$3 due to the price cut (i.e., \$1 for each of the three units previously sold at a price of \$6 each). The net change in revenue is \$2 (= 5 – 3). The implication then is that the MR of the fourth unit is below the price of the fourth unit. This point is indicated in Figure 9.5 where the MR_2 is equal to \$2 and the quantity of output is 4 units. Because the MR is generally less than the price of the product, we can draw the MR curve as a downward sloping curve that falls faster than the demand curve, as shown in Figure 9.5.

We have just seen that the MR curve for a monopolist falls faster than the demand curve facing the monopolist because the firm must reduce the price on all units sold when it wishes to sell an additional unit. It is helpful to contrast the MR curve facing the monopolist with the MR curve facing the perfectly competitive firm. If the perfectly competitive firm faces a constant market price of \$5 per unit, then we can calculate the firm's MR as shown in Figure 9.6.

Figure 9.6: The Marginal Revenue Curve Facing a Perfectly Competitive Firm



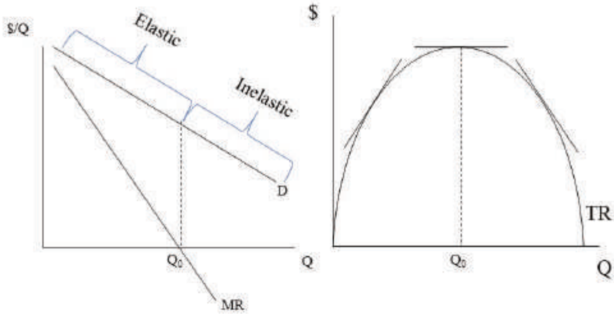
Using the information in Figure 9.6, we can see that the MR is calculated as follows:

$$MR = \frac{\Delta TR}{\Delta Q} = \frac{TR_2 - TR_1}{Q_2 - Q_1} = \frac{20-15}{4-3} = \$5 \; per \; unit$$

Without the need to reduce price to sell an additional unit, the perfectly competitive firm's MR is the same as the price charged.

We can also understand the shape of the MR curve facing the monopolist from a mathematical perspective if we consider the relationship between MR and TR. The MR is defined as ΔTR/ΔQ. If we consider the TR curve facing the monopolist, then we can see that the MR is nothing more than the slope of the TR curve. As Figure 9.7 shows, the slope of the TR curve at a given output level can be determined using the tangent line method.

Figure 9.7: The Relationship between the MR and TR Curves Facing a Monopolist



It should be clear that in the elastic portion of the demand curve, the MR is positive because the slope of the TR curve is positive. Similarly, at the peak of the TR curve where demand is unit elastic, the MR is equal to zero because the slope of the TR curve is zero. Finally, in the inelastic portion of the demand curve where the TR is falling, the MR must be negative because the slope of the TR curve is negative.

We can also infer at this stage that the monopolist will never produce where demand is inelastic. If the monopolist produces in this region of the demand curve then MR will be negative, which means that TR will be falling. A firm can never increase its economic profits by producing in this region because its TR will be falling even as its total cost (TC) is rising due to the increase in production. At this stage, we have not even incorporated TC into the analysis of the pure monopolist, but it is certain that higher production increases TC. As a result, economic profit must fall if the firm produces where demand is inelastic. Therefore, the monopolist will only produce where MR is greater than or equal to zero. This result tells us that a monopolist will tend to produce relatively less output rather than more, but at this stage, we cannot draw a more precise conclusion about the profit-maximizing choice of the monopolist.

Short Run Profit Maximization: Rules and Cases

Now that we have fully explored the revenue structure of the pure monopolist, we can turn to the question of profit-maximization. That is, which output level will be chosen and which price will be charged to maximize the firm's economic profit in the short run? To answer this question, we must introduce production cost into the analysis. We will assume that the cost structure facing the pure monopolist is identical to the cost structure elaborated in Chapter 7 and applied to the perfectly competitive firm in Chapter 8. It also turns out that the two rules of profit maximization in the short run that were applied in the case of a perfectly competitive firm are also applicable in the case of the purely monopolistic firm. That is:

1. *The firm should produce where MR=MC.*
2. *The firm should only produce a positive output level when P≥AVC.*

Table 9.2 provides detailed information facing a monopolistic firm.

Table 9.2: Demand and Cost Information for a Monopolistic Firm

Q	P	TR	MR	TC	ATC	MC	π
0	24	0	--	120	--	--	-120
10	22	220	22	200	20.0000	8	20
20	20	400	18	260	13.0000	6	140
30	18	540	14	300	10.0000	4	240
40	16	640	10	340	8.5000	4	300
50	14	700	6	400	8.0000	6	300
60	12	720	2	480	8.0000	8	240
70	10	700	-2	580	8.2857	10	120
80	8	640	-6	700	8.7500	12	-60
90	6	540	-10	840	9.3333	14	-300
100	4	400	-14	1000	10.0000	16	-600

The first two columns contain given information about the market demand schedule that the firm faces. Clearly, the market demand curve is downward sloping. TR is calculated as the product of price and quantity demanded in the market. MR is calculated in a manner just like that described earlier in this section. For example, we calculate the MR of the 20th unit of output as follows:

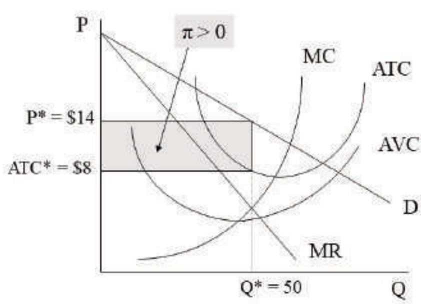
$$MR = \frac{\Delta TR}{\Delta Q} = \frac{TR_3 - TR_2}{Q_3 - Q_2} = \frac{400-220}{20-10} = \$18 \; per \; unit$$

The TC information in Table 9.2 is given, and the ATC, MC, and economic profit (π) calculations are all carried out as in earlier chapters.

A few details from Table 9.2 deserve special emphasis. The TR peaks at 60 units of output. At higher output levels, TR begins to fall and MR becomes negative. We can infer that demand becomes inelastic beyond 60 units of output. Because we are using *discrete* data for the level of output (e.g., 0, 10, 20, …), the MR is close to zero but not exactly equal to zero when the TR reaches its peak. As a result, we know that the firm must produce 60 units of output or less to maximize its economic profit since those output levels correspond to the elastic part of the demand curve. To be more precise, we can now apply the rule that the firm will produce where MR = MC (if P ≥ AVC). In the table, MR and MC are equal to $6 per unit at an output level of 50 units. We do not have specific information for AVC in the table, but since P > ATC at 50 units of output, P must necessarily exceed AVC. The firm earns its maximum economic profit of $300 at 50 units of output. It might be noticed that 40 units of output will also achieve this same amount of economic profit. Again, it is our use of discrete data that is responsible for this result, and so it is always best to focus on the rule that MR = MC.

Figure 9.8 shows how the monopolist's situation appears graphically.

Figure 9.8: Case 1 – The Monopolist Earning Economic Profits (P > ATC)



It is important to notice that even though the profit-maximizing output level is found directly below the MR = MC intersection, the profit-maximizing price that is charged is found above the intersection on the market demand curve. The area of the shaded box represents the firm's economic profit of $300, which may be calculated as the product of the amount by which P exceeds ATC (= $6 per unit) and the quantity sold (= 50 units). Alternatively, TR is $700 (= $14 per unit times 50 units) and TC is $400 (= $8 per unit times 50 units). The economic profit of $300 is simply the difference between the two amounts.

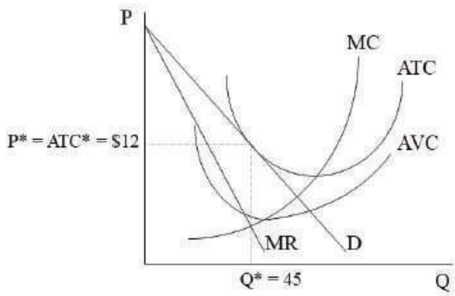
This analysis demonstrates that the goal of the monopolist is not to rip off the consumer. The monopolist aims to maximize economic profit. It could charge the consumer a higher price if it wished to do so, but a higher price would reduce its economic profit. We know that economic profit would be lost because at a higher price, the firm's MR exceeds its MC. Similarly, the firm could increase the quantity sold by reducing its price, but this action too would fail to maximize economic profit because a lower price would cause MC to exceed MR.

It is worth emphasizing that the monopolist does not have a supply curve like the perfectly competitive firm.[6] As we learned in Chapter 8, the perfectly competitive firm's profit-maximizing quantity changes as the market price changes for reasons beyond the firm's control. As the market price changes, we observe different quantities supplied by the perfectly competitive firm at each price. The monopolist, on the other hand, *chooses* the market price to maximize its economic profit. As a result, the price charged does not fluctuate for reasons beyond the monopolistic firm's control. Therefore, the only relevant combination of output and price for a monopolist is the one that it chooses to maximize its economic profit. It is not possible to trace out a series of combinations of price and quantity supplied in the case of the monopolist.

In addition, the economic profits that the monopolist earns may persist even in the long run due to the barriers to entry that prevent competitors from entering the market.[7] In the case of perfect competition, economic profits serve as a signal to firms outside the industry to enter. The lack of entry barriers leads to a rise in the market supply and a reduction in the market price until the economic profits are eliminated. No such elimination of the economic profits can occur in the case of pure monopoly. Although the monopolist is protected from competition, its economic profits are not guaranteed. If consumer demand falls then the market demand curve will shift to the left, which may eliminate the economic profits.[8] Similarly, a rise in production costs may occur, which would cause the ATC curve to shift upwards.[9] The result again would be a reduction in economic profits.

These possibilities lead us to consider additional scenarios that may face the monopolist in the short run. Figure 9.9 represents the case where a monopolist is operating at the break-even point.

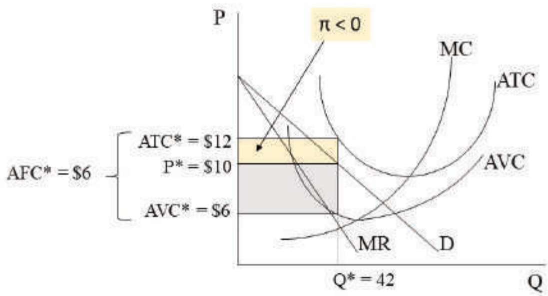
Figure 9.9: Case 2 – The Monopolist at the Break-Even Point (P = ATC)



In Figure 9.9, the firm maximizes its economic profit by producing 45 units of output. This output level corresponds to the MR = MC intersection, and the price of $12 per unit at this output level is clearly above AVC. In fact, at Q = 45 units, the price equals the ATC. As a result, the firm's TR is $540 (= $12 per unit times 45 units) and the firm's TC is $540 (= $12 per unit times 45 units). Its economic profit is, therefore, equal to zero and so it is not represented as a shaded area in the graph. It should be recalled that this situation is acceptable to the monopolist because the firm is covering all its costs, including the opportunity cost of producing in this industry. Greater economic profits are always desired, of course.

Figure 9.10 shows another scenario in which the monopolist maximizes its economic profit by producing 42 units of output.

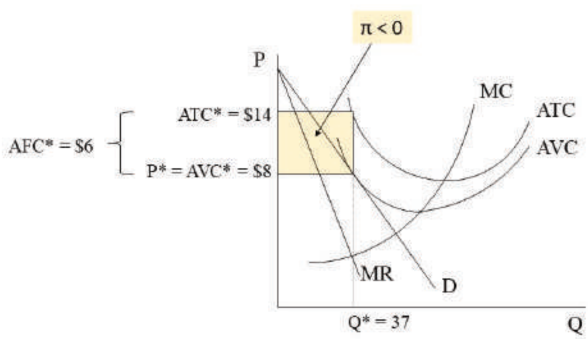
Figure 9.10: Case 3 – The Monopolist Operating at a Loss (ATC > P > AVC)



At that output level, MR = MC but also the profit-maximizing price of $10 is above the AVC of $6. The firm's TR in this case is $420 (= $10 per unit times 42 units), and the firm's TC is equal to $504 (= $12 per unit times 42 units). The economic profit then is equal to -$84 (= $420 – $504), which is also equal to the area of the top shaded box in Figure 9.10. If the firm is suffering a loss as indicated by its negative profit, then why does the firm operate in the short run? The reason is that the loss from shutting down would be even larger. As we discussed in Chapter 8, whenever a firm shuts down in the short run, its economic profit is equal to –TFC. The TFC can be calculated as the product of the AFC and the level of output. In this case, the TFC is equal to $252 (= $6 per unit times 42 units), which is also equal to the combined area of the two shaded regions in Figure 9.10. Clearly, an economic profit of -$252 is much worse for the firm than an economic profit of -$84. Therefore, the monopolist will choose to operate in the short run at a loss.

The next case we will consider is one in which the monopolist maximizes its economic profit by producing 37 units of output as shown in Figure 9.11.

Figure 9.11: Case 4 – The Monopolist at the Shut Down Point (P = AVC)

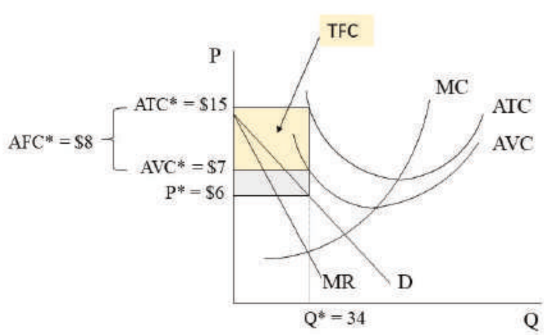


In this case, the firm charges a profit-maximizing price of $8 per unit, which is also equal to the firm's AVC at this output level. Because the shutdown rule states that the firm will produce where MR = MC if P is greater than or equal to AVC, we conclude that the firm will operate in this case. The firm's TR in this scenario is equal to $296 (= $8 per unit times 37 units). The firm's TC is equal to $518 (= $14 per unit times 37 units). Therefore, the economic profit is equal to -$222. In addition, the firm's TFC is equal to $222, which may be calculated as the product of AFC (= $6) and Q (= 37 units). Therefore, if the firm shuts down, it will suffer an economic loss of $222 (= -TFC), which is exactly equal to its loss from operating. In Figure 9.11, the shaded region represents both the TFC and the economic loss from operating. Even though the firm is indifferent between operating and shutting down, we conclude that the firm operates.

Another way to argue that the firm is indifferent between operating and shutting down is to note that the firm is earning just enough revenue to cover its TVC. The firm's TVC is equal to $296 because TVC may be calculated as the product of AVC (= $8) and Q (= 37 units). In other words, the firm can just pay wages out of revenue so it is only the TFC that it cannot cover when it operates.

The final case that we will consider for the short run is the case in which the monopolist decides to shut down as shown in Figure 9.12.

Figure 9.12: Case 5 – The Monopolist Shuts Down in the Short Run (P < AVC)



In this case, the firm produces no output because the price of $6 per unit, which is directly above the MR = MC intersection, is below the AVC of $7 per unit. Because the monopolist shuts down, its TR is \$0 and its TVC is \$0. It must still pay its TFC and so its economic profit is equal to –TFC. As always, the TFC may be calculated as the product of AFC and Q, which is here equal to \$272 (= \$8 per unit times 34 units). The firm's economic loss of \$272 is depicted as the top shaded box in Figure 9.12. If the firm were to produce 34 units and charge a price of \$6 per unit, then its economic loss would be much greater. In that case, the monopolist's TR would be \$204 (= \$6 per unit times 34 units), its TC would be \$510 (= \$15 per unit times 34 units), and its economic profit would be -\$306 (= \$204 – \$510). The economic loss in that case would be equal to the combined area of the two shaded boxes in Figure 9.12. As a result, the firm opts to shut down and limit its losses to the TFC.

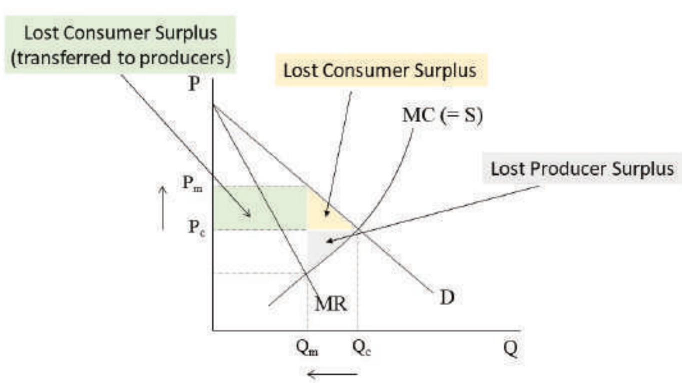
Implications for Efficiency and the Long Run

The five cases that we considered in the previous section are all possible scenarios that the monopolist will face in the short run. As we observed, it is only in cases 1-4 that the monopolist will operate in the short run. Whenever price falls below AVC, as in case 5, the firm shuts down. The reader might wonder which of these cases is possible in the long run. In the long run, the firm will withdraw its capital and exit the industry if it is experiencing economic losses. Cases 3-5 all include economic losses and so these cases cannot apply to the monopolist in the long run. On the other hand, cases 1 and 2 involve an economic profit that is greater than or equal to zero. As we have seen, an economic profit of zero will be acceptable even to a monopolist in the long run because all costs are being covered, including opportunity costs. Furthermore, we have seen that barriers to entry make long run economic profits a real possibility because competitors cannot enter and drive down the price. Therefore, either of these two cases is a possible long run equilibrium.

Our next task is to consider the efficiency implications of monopoly markets. We have seen that perfectly competitive markets achieve economic efficiency in the sense that they lead to least-cost production (productive efficiency) and the appropriate quantity of the good produced (allocative efficiency). They also maximize the total of consumer surplus and producer surplus. It is worth asking whether monopoly markets achieve the same desirable outcomes.

Figure 9.13 allows us to compare the efficiency effects for the cases of pure monopoly and perfect competition.

Figure 9.13: The Harberger Triangle and the Rectangle of Redistribution in Monopolistic Markets



In Chapter 8, we learned that the marginal cost (MC) curve of a perfectly competitive firm is the same as the perfectly competitive firm's supply curve (for MC above minimum AVC). It followed that the aggregation of the MC curves of all the perfectly competitive firms in the industry would yield the market supply curve. If we assume that the monopolist operates with the same production technology as the perfectly competitive firms, then we can assume that the monopolist's MC curve is identical to the market supply curve that would emerge if the market was perfectly competitive. For this reason, the monopolist's MC curve in Figure 9.13 is also labeled as a perfectly competitive market supply curve (= S). This approach allows us to compare the efficiency effects of the two market structures on the same graph.

If the market is perfectly competitive, then the equilibrium outcome is given by the intersection of market supply and market demand. The equilibrium price in that case would be P_c and the equilibrium quantity would be Q_c in Figure 9.13. If the market is monopolistic, on the other hand, then the monopolist produces Q_m where MR and MC intersect and charges a price of P_m as shown on the market demand curve above that point. Clearly, the monopolist produces a smaller quantity and sells it at a higher price than would emerge in a perfectly competitive market. The failure to produce where marginal benefit (represented by demand) and marginal cost intersect indicates that the monopoly market does not achieve allocative efficiency. Indeed, this result is the main reason that neoclassical economists object to monopoly markets. Unlike consumers who tend to complain about high monopoly prices, neoclassical economists tend to focus much more on the fact that monopolies do not produce as much of the product as would be produced in a perfectly competitive market.

The monopoly firm also fails to achieve least-cost production. If we look back at Figure 9.8, for example, in which the firm is enjoying economic profits, we can see that the price charged is clearly above the minimum ATC, which exists at the intersection of MC and ATC. If the firm was to produce more, then its ATC would fall due to further specialization within its production plant. It chooses not to expand production, however, because holding back production allows the firm to raise the price by more than enough to make up for the higher unit cost.

Returning to Figure 9.13, we can also see that the reduction in output that occurs in the monopoly market leads to a loss of consumer surplus and producer surplus. The reader should recall that the consumer surplus is equal to the area above the current price and below the demand curve. In this case, the increase in price that the monopolist pursues causes the consumer surplus to shrink by the amount of the shaded rectangle and the top shaded triangle. The top shaded triangle, however, represents a pure loss to society because it corresponds to the reduction in quantity that accompanies the monopolization of the market. These units are not produced anymore and so this consumer surplus is completely lost. The shaded rectangle, however, corresponds to units that are still produced in the monopoly market. This area represents a transfer to the producers in the form of producer surplus. For this reason, the shaded rectangle is referred to as the **rectangle of redistribution**. These units are still sold but now at the higher monopoly price.

The reader should recall that producer surplus is represented as the area below the current price and above the supply curve (or the marginal cost curve). The reduction in output that occurs with the monopolization of the market thus leads to the loss of the bottom shaded triangle in Figure 9.13. Because these units of output are no longer produced, this producer surplus is simply lost. This loss of producer surplus, of course, raises the question as to why the monopolist would restrict output when it leads to a loss of producer surplus. The reason is that this reduction in output makes possible a rise in price, which allows the monopolist to capture a significant part of the consumer surplus. A net gain for the monopolist is the result.

What happens in this case is that the total surplus (TS) that society realizes shrinks due to the loss of consumer surplus and producer surplus. This loss of TS represents an efficiency loss from the monopolization of the market. The two shaded triangles in the graph represent the deadweight loss of monopoly. The larger triangle formed from the combination of these two smaller triangles is called a **Harberger Triangle** after the American economist Arnold Harberger for whom it is named. In summary, the total economic pie shrinks due to the monopolization of the market even as the monopolist grabs a larger share for itself at the expense of consumers.

Additional characteristics of monopoly markets also strengthen the case that such markets tend to be inefficient. Specifically, monopolistic firms have an incentive to incur additional costs to maintain the barriers to entry that keep competitors out of their markets. Firms may hire lobbyists to pressure lawmakers to extend their patents and copyright protection. Lawyers may be hired to initiate lawsuits against those believed to be infringing on the firm's rights of ownership of patents and copyrights. Because these expenses are incurred to protect the firm's economic profits rather than to increase production, they are regarded as wasteful from a social perspective. Neoclassical economists label these efforts to appropriate benefits that exceed the economic cost of the required resources as **rent-seeking activities**. To the extent that these legal barriers to entry encourage innovation, that benefit may provide some economic justification for these expenses. Any costs that exceed what is needed to encourage innovation, however, are regarded as socially wasteful in neoclassical theory.

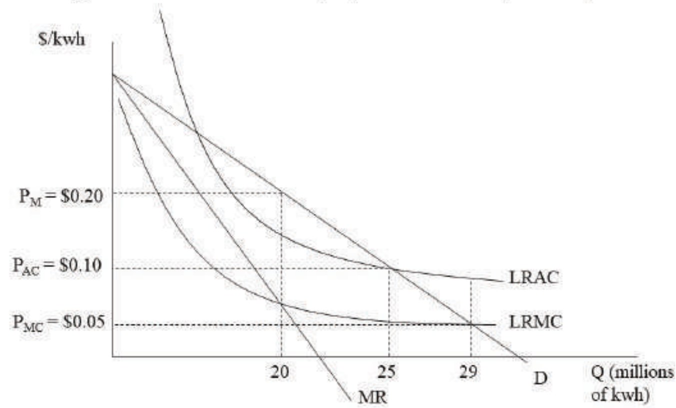
In terms of long run economic growth, monopolies are also often accused of being technologically backwards.[10] That is, because they are not subject to competitive pressure, the claim is made that they will not have the incentive to innovate and keep production costs down in the same way that a perfectly competitive firm would. An example of a large firm that lost considerable market share due to a failure to innovate is General Motors. Although it was not a pure monopoly in the American automobile market given the presence of other large automakers, it was the market leader in the mid-twentieth century. Its failure to keep up with market trends caused it to lose considerable market share to Japanese auto producers as Japanese auto companies began selling smaller, more fuel-efficient cars in response to the rise in gas prices in the 1970s. Another interesting example of the tendency of large corporations to stagnate technologically is U.S. Steel's refusal to develop a new steel beam known as a Grey beam at the beginning of the twentieth century. By 1926, U.S. Steel was secretly producing the Grey beam in violation of Bethlehem Steel's patent rights. Eventually, U.S. Steel was forced to pay royalties to Bethlehem Steel in exchange for permission to produce the Grey beam.[11]

Natural Monopoly and Price Discriminating Monopoly

Overall, monopolies appear to be economically inefficient. Nevertheless, we can identify some exceptions to this general rule. One example is when a natural monopoly exists. As we have seen, natural monopolists enjoy an economic cost advantage due to economies of scale. It is this cost advantage that allows them to dominate the market. If significant economies of scale exist in an industry, then our assumption breaks down that the MC of the monopolist is the same as the market supply curve in a perfectly competitive market. With significantly lower per unit production costs, the natural monopolist may end up being *more* efficient than the perfectly competitive market.

As we discussed earlier in this chapter, a natural monopolist enjoys falling LRAC over a large range of output. Public utilities are frequently cited examples of natural monopolies. The startup costs are very large and so these costs are spread out over a large quantity of output. Most areas are served by a single electric company, a single natural gas company, and a single water company. Imagine the inefficiencies that would arise if many suppliers each had their own electrical wires, gas lines, and water delivery systems. Consider the case of an electric company that enjoys large economies of scale in production and so monopolizes the local electricity market. This case is depicted in Figure 9.14.

Figure 9.14: A Natural Monopoly in the Electricity Industry



In Figure 9.14, the quantity of output is measured in millions of kilowatt hours (kwh) used per month. The price and unit costs are measured in dollars per kilowatt hour. Economies of scale are clearly present as can be seen from the declining LRMC and LRAC curves. The monopolist also faces a downward sloping demand curve and has a corresponding MR curve that declines more quickly than price. If the monopolist is allowed to operate freely, it will maximize its economic profit by selling 20 million kwh at a price of $0.20 per kwh. Because the monopoly price (P_M) exceeds LRAC at 20 million kwh, clearly the monopolist will earn an economic profit.

Even though the natural monopolist has an economic cost advantage, its desire to maximize its economic profit might prevent the firm from operating as efficiently as possible from a social perspective. For this reason, arguments in favor of public regulation of privately owned monopolies are sometimes made. If regulation is chosen, then a state regulatory board would most likely require the natural monopolist to set its price below the profit-maximizing price of $0.20 per kwh. Two possibilities are depicted in Figure 9.14. One possibility is to require the electricity provider to charge a price equal to LRMC ($0.05 per kwh in this example) as is the case in a perfectly competitive market in long run equilibrium. This regulatory strategy is known as **marginal cost pricing**. If the firm is also required to produce the perfectly competitive level of output of 29 million kwh, then the price is much lower and the quantity sold is much higher than in the profit-maximizing outcome. The problem is that the regulatory board has forced the monopolist to suffer losses as indicated by the fact that the price charged is below LRAC at that output level.

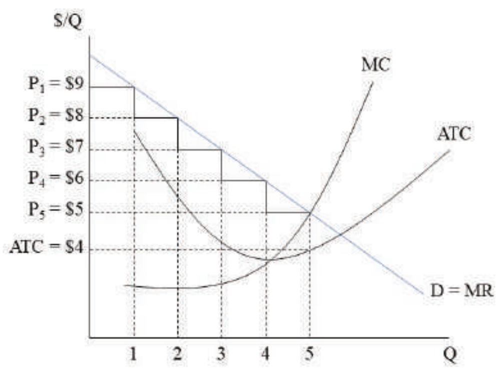
An alternative regulatory strategy is to require the monopolist to charge a price equal to LRAC. This regulatory strategy is known as **average cost pricing** and guarantees that the firm will earn an economic profit of zero. If the firm charges a price of $0.10 per kwh and produces an output level of 25 million kwh, then the firm will break even. This pricing rule leads to a larger quantity of output sold than in the monopoly outcome but not as much as the marginal cost pricing rule. It also leads to a lower price than the monopoly price but one that is higher than the perfectly competitive price. It thus represents a compromise between the two possible outcomes that is still acceptable to the natural monopolist, even in the long run since economic profits are zero.

Another possible scenario in which a monopolist may be more efficient than the basic model occurs when the monopolist can engage in **price discrimination**. Price discrimination refers to the practice of charging different prices to different consumers of the same product. For example, student discounts on restaurant meals or movie tickets are good examples of price discrimination. Regular customers must pay full price but the students are given a discount to encourage them to patronize specific establishments. Senior discounts on restaurant meals and kids' menus are other examples. Additional examples include bulk discounts. When a consumer buys a 24-pack of soda, for example, she will generally pay a lower price per can than if she purchases a 6-pack of soda.

The examples of price discrimination cited above involve different groups of consumers being charged different prices based on some distinguishing characteristic of the consumer or the quantity that is purchased. The purest form of price discrimination, however, occurs when every consumer in a market is charged the maximum price she is willing and able to pay. This form of price discrimination is referred to as **perfect price discrimination** because it is only possible when the seller can perfectly discriminate between consumers. For this form of price discrimination to occur, the firm must be a monopolist. If competitors can charge lower prices than the maximum that the consumers are willing and able to pay, then this pricing strategy will be undermined very quickly. The monopolist must also have complete knowledge of the market demand curve. Otherwise, it will be impossible to identify the maximum prices that consumers are willing and able to pay. Finally, it must be impossible for consumers to resell the units they purchase.[12] If a consumer who is willing and able to pay a low price can turn around and sell that unit to a consumer who is willing and able to pay a much higher price, then the monopolist may fail to make the sale to the consumer who is willing and able to pay the high price, particularly if the consumer who resells the unit charges a slightly lower price for that unit.

Figure 9.15 shows the case of a monopolist that is capable of perfect price discrimination.[13]

Figure 9.15: Perfect Price Discrimination



The major difference between this graph and the graph of the single pricing monopolist is that this monopolist's MR curve is identical to the demand curve facing the firm. With a single pricing monopolist, the MR curve is steeper than the demand curve because the firm must cut the price on all units sold to sell another unit. When the monopolist can perfectly price discriminate, however, it does not need to cut the price on other units sold to sell

one additional unit. As a result, the MR of an additional unit is the same as the price. Therefore, the market demand curve and the MR curve are the same in the case of perfect price discrimination. In Figure 9.15, the first unit is sold for $9 per unit, which is the maximum price that a consumer is willing and able to pay for that unit. The second unit sells for $8. Notice that the first unit still sells for $9 even though the second unit sells for $8 because perfect price discrimination occurs in this case. The MR of the second unit then is $8 because TR rises by the full amount of the price of that unit. The third, fourth, and fifth units sell for $7, $6, and $5, respectively. Notice that the monopolist will stop producing at 5 units because at this point MR = MC. The firm's TR is calculated as the sum of the different prices charged as follows:
$TR = 9 + 8 + 7 + 6 + 5 = \$35$

The monopolist's total cost is calculated as it always is. If we multiply the ATC at the profit-maximizing output by the quantity produced, then the TC is equal to $20 (= $4 per unit times 5 units). The firm's economic profit then is equal to $15 (= $35 – $20).

Because the perfectly price discriminating monopolist can charge the maximum price the consumer is willing and able to pay, the firm appropriates the consumer surplus that consumers normally enjoy when they pay the single market price. This transfer to the monopolist allows the monopolist to increase its economic profit at the expense of the consumer. This result is important because it demonstrates that price discrimination is not a practice that is adopted to help special groups like senior citizens, parents with small children, and college students. The practice is adopted because it maximizes economic profits. By charging lower prices to senior citizens and kids and higher prices to everyone else, a restaurant can appropriate (at least in part) the smaller consumer surpluses of the senior citizens and children and the larger consumer surpluses of everyone else.

We can also see that a perfectly price discriminating monopolist is more efficient than a single pricing monopolist. Specifically, if we assume once more that the MC curve of the monopolist is the same as the market supply curve if the market was instead perfectly competitive, then we can evaluate the efficiency effects of price discriminating monopoly relative to perfect competition. In Figure 9.15, the perfectly competitive equilibrium can be found at the intersection of the market demand curve and the MC curve of the monopolist (i.e., the perfectly competitive supply curve). The perfectly competitive output level is 5 units of output, which is the same as the profit-maximizing output of the price discriminating monopolist. In other words, the price discriminating monopolist achieves allocative efficiency. It fails, however, to achieve least-cost production because the ATC of $4 per unit is above the minimum ATC, which occurs at the intersection of the ATC and MC curves. The implication is that perfect price discrimination does not shrink the economic pie, but it does redistribute a large part of the pie away from consumers and towards the monopolist.

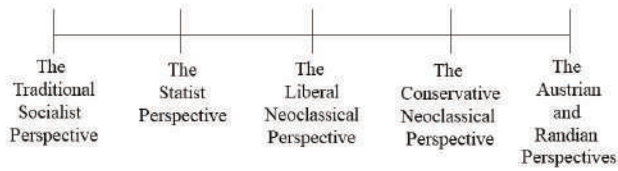
Government Policy Regarding Monopoly

Because monopolists tend to operate inefficiently, the question arises as to how society should respond when monopolistic markets emerge. We can identify five possible responses. Each response may be associated with a specific school or schools of economic thought.

1. The Austrian and Randian perspectives: Allow private monopolies to thrive without government intervention.
2. The conservative neoclassical perspective: Divide inefficient monopolies into competing firms by enforcing antitrust laws.
3. The liberal neoclassical perspective: Publicly regulate privately owned natural monopolies and divide monopolies into competing firms with even stricter enforcement of antitrust laws.
4. The statist perspective: Create public monopolies through nationalization (state acquisition)
5. The traditional socialist perspective: Grant the power to bodies of working people to cooperatively plan production in monopoly firms in coordination with other bodies of working people and consumers.

We can place each of these views along the political spectrum ranging from far right to far left as shown in Figure 9.16.

Figure 9.16: The Range of Government Policies Regarding Monopoly



We will briefly discuss each of these responses beginning with the neoclassical perspectives since we have already explored the neoclassical theory of monopoly in detail. According to conservative neoclassical economists, the ideal market structure is the perfectly competitive one. Therefore, efficiency can be enhanced by breaking up monopolies into competing firms, effectively transforming a monopolistic market structure into a perfectly competitive market structure. The notion that government intervention can be used to create a more competitive environment is one that has a long history in the United States. The 1890 Sherman Act was the first major piece of federal legislation to outlaw unreasonable restraints of trade. In the early twentieth century, federal action was taken against Standard Oil, U.S. Steel, American Tobacco, and other firms because they had allegedly established illegal monopolies in their industries. The Clayton Act of 1914 and the creation of the Federal Trade Commission (FTC) that same year to enforce the nation's antitrust laws further strengthened this instrument of federal control over markets. The conservative neoclassical perspective takes seriously the benefits of perfect competition, but its proponents prefer that this federal tool be used conservatively (i.e., only when a clear and convincing case can be made that a monopoly is firmly established in a specific industry and is harming consumers).

The liberal neoclassical perspective advocates even stricter enforcement of the nation's antitrust laws as well as public regulation of privately owned natural monopolies. Again, the aim is to force the natural monopolist to produce a larger output and charge a lower price than its profit-maximizing choices of these variables. In theory, successful regulation using average cost pricing or marginal cost pricing rules will move the industry closer to the perfectly competitive outcome. Conservative neoclassical economists tend to object to regulation because regulatory boards are susceptible to **regulatory capture**.[14] That is, these boards may come under the influence of the monopolist itself, which can use the coercive power of government to impose its will, thereby creating an even worse situation than that which exists in an unregulated environment. In addition to their advocacy of public regulation of monopolies, liberal neoclassical economists have a lower threshold at which point they advocate antitrust action against potential violators. The remaining perspectives on responses to monopoly are discussed in the next two sections.

The Austrian and Randian Critiques of Neoclassical Monopoly Theory

The Austrian and Randian perspectives of monopoly theory and antitrust laws are similar and so have been treated together. Differences do exist among adherents of these worldviews, but their commonalities are much greater than the differences. The former chairman of the Federal Reserve Board, Alan Greenspan, was once a member of Ayn Rand's inner circle. His perspective on U.S. antitrust laws represents the Randian perspective quite well.[15] Specifically, Greenspan challenges the conventional view of the western railroad monopolies that operated in the late nineteenth century. The western railroads have been accused of operating independently of competition even though many were poorly built. Greenspan argues, however, that their monopoly power derived from government subsidies and government restrictions.

According to this Randian perspective, if a single firm becomes dominant in an industry, that fact alone is not reason enough to oppose its existence. Assuming its dominance is the result of its own efficient operation and competitors are free to enter the industry and compete, then the monopolist should not be punished for its success. The only objectionable monopoly, according to this view, is a **coercive monopoly**. A coercive monopoly is one that maintains its dominance due to legal restrictions that prevent competitors from entering. Once such restrictions are removed, a free capital market will ensure that competitors will enter the industry if they expect that they can compete with the monopolist, argues Greenspan. If the monopolist remains the dominant firm even in the presence of a free capital market, then it will not necessarily have its profits reduced. If the firm is highly efficient and keeps production costs low, then its profit margins may be large. Greenspan argues that a free capital market only ensures that a monopolist that earns big profits by charging high prices will soon face competition that drives those profits down. He identifies ALCOA as an example of a monopoly firm that maintained its monopoly in the market for aluminum for many years prior to World War II due to cost-cutting and efficiency rather than high prices. This opinion contrasts with our earlier suggestion that ALCOA's control of a key resource was the source of its market power. In summary, a monopolist should never be punished when it achieves and maintains its dominance in an environment of free competition and voluntary exchange.

The Austrian critique of neoclassical monopoly theory resembles the Randian critique. D.T. Armentano, for example, explains that neoclassical economists have placed great emphasis on non-legal barriers to entry like economies of scale and commercially successful product differentiation as sources of monopoly power.[16] Examples of product differentiation include the annual models of automobile companies. Armentano explains that consumers clearly prefer to pay the higher prices that accompany such differentiation. Similarly, consumers could pay the higher prices of firms that do not enjoy scale economies if they wished to do so. From the consumer's perspective, resources *are* allocated efficiently and so to use antitrust laws against such firms is in complete opposition to the wills of consumers. Armentano also warns against evaluating the efficiency of monopoly markets against the neoclassical model of perfect competition. He explains that the perfectly competitive model is completely static (i.e., devoid of time and process). As a result, all the features of actual competition, such as product differentiation, advertising, price discrimination, and innovation, are treated as monopolistic practices that lead to the misallocation of resources. If we regard all such activities as monopolistic, then most competitive interaction will be viewed as requiring government intervention, in Armentano's view.

Along similar lines, Armentano explains how the Austrian economist Murray Rothbard argues that it is not possible to conceptually distinguish between a monopoly price and a free market competitive price. If we compare the price of a good in a competitive market with the price of a good in a monopolistic market, then we are simply comparing the prices of two different goods. If consumers are willing to pay a higher price for one of the goods, then the price paid is consistent with the supply and demand for that good just as the lower price is consistent with the supply and demand for that good. Additionally, if the market price of a good rises over time, why should we assume that the initial price is the competitive price and the new price is a monopolistic one? All free market prices ultimately emerge from the interaction of supply and demand and attempts to label some prices as monopoly prices and other prices as competitive prices are simply efforts to justify government interference in the free market.

The Marxian Theory of Monopoly Capital

Marxian economists have also constructed analyses of **monopoly capital**. These analyses have been complicated, however, by the fact that Marx primarily investigated the conditions of competitive capitalism. Nevertheless, he did recognize the tendency towards monopoly within capitalist industries. One source of this tendency, Marx refers to as the **concentration of capital**. As an individual enterprise accumulates capital over time, it grows as it devotes a portion of its profits to expansion. The **centralization of capital**, on the other hand, refers to the tendency of strong, productive enterprises to absorb weaker, less productive enterprises and may be an important factor during capitalist crises. The process is never complete, however, as competition between large capitalists inevitably breaks out again. In Marx's view, competition gives rise to monopoly and monopoly gives rise to competition.

When Marxian economists have referred to "monopoly capital," the phrase has been used more generally than the strict neoclassical definition of a single seller in a market for which no close substitutes exist and in which barriers to entry keep competitors from entering. Instead, it seems to refer to industries in which the concentration and centralization of capital have become so great that the law of value no longer prevails. That is, a monopoly capitalist must have the power to increase the price of a commodity above the value that would prevail if the market was intensely competitive. In other words, the monopoly price will be higher than the competitive price that would be charged if the value of the commodity was determined by the SNALT embodied in it. Marxian economists have not put forward an explanation of any economic law that would determine the magnitude of the difference between the monopoly price and the competitive value.

Because of the difficulties associated with the identification of a precise economic law to explain the difference between monopoly price and competitive value, Marxian economists have concentrated more on general qualitative analyses of monopoly capital. The most famous of these analyses is Paul Baran and Paul Sweezy's *Monopoly Capital* (1966). Baran and Sweezy argue that over time within capitalist societies a persistent rise in the **economic surplus** occurs, which is simply the difference between the total social product and the cost of producing it. Because of income inequality, insufficient aggregate demand exists to absorb the surplus. Capitalist consumption and investment are not sufficient to absorb the surplus, and so stagnation is the expected outcome of monopoly capitalism. Some counteracting forces do exist, however, which include the **sales effort**. The sales effort refers to advertising and product differentiation, which may stimulate demand to a certain degree but which mainly create an enormous amount of waste. Government expenditure is the other major counteracting factor that Baran and Sweezy discuss. Because capitalists have a major influence over government officials, however, they emphasize military expenditures much more than spending on social programs and public assistance because the latter might threaten the class structure of society.[17]

In 2006, John Bellamy Foster and Fred Magdoff argued that the stagnation problem that Baran and Sweezy identified as consistent with the monopoly capital phase had worsened. While the counteracting factors are "superficial, weaker, and self-limiting," the stagnation tendency is "deeply rooted, powerful, and persistent." At the same time, Foster and Magdoff argue that the capitalist system has found new ways to offset (at least partly) the stagnation tendency of monopoly capital. They explain that the explosive growth of finance has led to a new hybrid phase of the capitalist system that they refer to as **monopoly-finance capital**. They explain that the massive growth of finance has led to new outlets for the economic surplus in the finance, insurance, and real estate (FIRE) sectors in the form of new investment in office buildings and equipment. This counteracting factor was limited, however, in that most money capital has been used for speculation in the financial markets rather than for real investment in productive capacity. Due to the inconsistent connection between the financial sector and the industrial sector, the financial explosion has been unable to compensate for the stagnation tendency of monopoly capital enough to revive the rapid accumulation of capital.[18]

Following the Economic News

The major producer of beer in South Africa, South African Breweries (SAB), had a market share of around 98 percent in 2007.[19] For all practical purposes, SAB had acquired a monopoly in the South African beer market. The firm is engaged in price discrimination with different prices charged in different regions as well as between urban and rural areas. The South African Competition Commission investigates restrictive business practices within the country. According to the Competition Commission's economic expert Simon Roberts, who has analyzed company documents, the firm's pricing strategy has involved "determining what the market can bear." According to Roberts, SAB was charging the highest prices it could in some areas with the hope of balancing out other areas. In other words, the firm has tried to charge the maximum that consumers are willing and able to pay, which is perfectly consistent with price discrimination. A major problem that SAB encountered, however, with this pricing strategy is that the pricing of beer was so different across regions that it gave rise to so-called "runners." That is, some distributors would purchase beer from cheaper regions and resell them in more expensive regions to make a profit. As we learned earlier in this chapter, resale must be impossible if perfect price discrimination is to occur. Although the case involving SAB is not a case of *perfect* price discrimination, clearly the reselling of the beer in different regions has interfered with the company's strategy of imperfect price discrimination. Due to complaints that liquor wholesalers and retailers brought to the Competition Commission in 2004, a special tribunal agreed to hear the complaints until August 16, 2013.[20]

Summary of Key Points

1. A monopolist is a single seller in a market for which no close substitutes exist and into which competitors cannot enter due to barriers to entry.
2. Barriers to entry include copyrights, patents, licenses, ownership of key inputs, and economies of scale.
3. For a single-pricing monopolist, MR falls faster than price because the firm must cut the price for all units sold to sell another unit.
4. The profit-maximizing output for the monopolist occurs where MR = MC, and the price charged is set above this output level on the market demand curve.
5. The monopolist will operate in the short run if the price is at least as great as AVC. The firm will

operate in the long run if the price is at least as great as ATC.

6. Single pricing monopolists fail to achieve allocative efficiency and productive efficiency.
7. Natural monopolies and perfectly price discriminating monopolies can offset some of the inefficiencies of single pricing monopolies.
8. The possible government policies to address monopolization range from laissez-faire to outright nationalization or support for worker control.
9. The only kind of monopoly that is objectionable according to Austrian and Randian thinkers is one that is the product of government entry barriers (i.e., a coercive monopoly).
10. According to Marxian economists, stagnation is the dominant tendency in the monopoly phase of capitalism, but it is mitigated by the sales effort, government spending, and the expansion of finance.

List of Key Terms

Monopolies

Price-setter

Natural monopoly

Rectangle of redistribution

Harberger triangle

Rent-seeking activities

Marginal cost pricing

Average cost pricing

Price discrimination

Perfect price discrimination

Regulatory capture

Coercive monopoly

Monopoly capital

Concentration of capital

Centralization of capital

Economic surplus

Sales effort

Monopoly-finance capital

Problems for Review

1.Suppose a monopolist reduces the price of its product from $8 per unit to $5 per unit. The quantity demanded of its product subsequently rises from 200 units to 350 units. Calculate the marginal revenue. (Hint: Strictly apply the definition of marginal revenue in this case.) Is the firm operating in the elastic or inelastic portion of the linear demand curve? How do you know?

2. The first two columns in the table below represent the demand schedule facing a pure monopolist. Complete the rest of the table using MS Excel.

Problems for Review: Problem 2

Q	P	TR	MR	TC	ATC	MC	π
0	70		--	25	--	--	
1	66			45			
2	62			60			
3	58			70			
4	54			85			
5	50			105			
6	46			130			
7	42			170			
8	38			220			
9	34			280			
10	30			360			

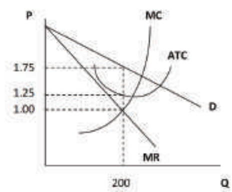
3. At which output level and price will this monopolist maximize economic profits? How do you know?

4. Create a single graph in MS Excel using the data above that shows the demand curve (D), the marginal revenue curve (MR), the marginal cost curve (MC), and the average total cost curve (ATC). Be sure to label both axes.

5. Using the graph below, identify the following, assuming the firm maximizes its economic profit in the short run:

- Output
- Price
- Total revenue
- Total cost
- Economic profit
- Will the firm operate in the short run?

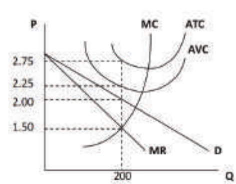
Problem #5



6. Using the graph below, identify the following, assuming the firm maximizes its economic profit in the short run:

- Output
- Total revenue
- Total cost
- Economic profit
- Will the firm operate in the short run?
- What is the firm's total fixed cost?

Problem #6



Notes

1. Bagdikian (2004), p. 72.
2. Ibid. Pp. 72.
3. Rys, Dan. "A Brief History of the Ownership of the Beatles Catalog." Billboard. January 20, 2017. Web. Accessed on May 7, 2018. https://www.billboard.com/articles/columns/rock/7662519/beatles-catalog-paul-mccartney-brief-history-ownership
4. Ibid.
5. This example is a commonly cited example in neoclassical textbooks. See Hubbard and O'Brien (2019), p. 512.
6. McConnell and Brue (2008), p. 430, offer a graphical method of demonstrating this point.
7. Bade and Parkin (2013), p. 407, emphasize that a monopolist's economic profits may persist indefinitely.
8. McConnell and Brue (2008), p. 431-432, mention that a demand reduction can threaten monopoly profits.
9. Bade and Parkin (2013), p. 407, consider an example of a rise in fixed cost that may cause monopoly losses.
10. Hubbard and O'Brien (2019), p. 522, point out that disagreement on this issue exists with some economists, like Joseph Schumpeter, arguing that firms with greater market power have the resources to test out new products in the marketplace and other economists arguing that the greatest innovation originates with small firms.
11. Hessen (1975), p. 267-269.
12. McConnell and Brue (2008), p. 437, identify the required conditions for perfect price discrimination.
13. This example is not a pure case of perfect price discrimination

because of the use of discrete quantities. It does represent a close approximation though.

14. See Chiang and Stone (2014), p. 232, for a discussion of George Stigler's views on regulation.
15. Greenspan (1967), p. 63-72.
16. Armentano (1978), p. 94-110.
17. A more complete description of Baran and Sweezy's book may be found in Hunt and Lautzenheiser (2011), p. 530-532.
18. Foster and Magdoff (2009), p. 63-67.
19. Cable, Jonathan. "SABMiller Says Heineken Calls Time on Amstel Licence." *Reuters*: London, March 13, 2007.
20. Most of the information in this section is from the following source: Benjamin, Chantelle. "SAB Abused its Monopoly, Says Economic Expert." *Mail & Guardian*: July 22, 2013.

Chapter 10

THEORIES OF IMPERFECTLY COMPETITIVE MARKETS

Goals and Objectives:

In this chapter, we will do the following:

1. *Construct* two key measures of industry concentration
2. *Identify* the key characteristics of monopolistic competition in neoclassical theory
3. *Explain* how monopolistically competitive firms behave in the short run and long run
4. *Describe* the social consequences of monopolistic competition
5. *Outline* the key characteristics of oligopoly in neoclassical theory
6. *Explore* several neoclassical oligopoly models
7. *Investigate* the game theoretic concepts of dominant equilibrium and Nash equilibrium
8. *Examine* a Post-Keynesian markup pricing model

Two Measures of Industry Concentration

In this chapter we will explore imperfectly competitive markets. These market structures fall in between the neoclassical ideal of perfect competition and its antithesis of pure monopoly. The number of firms in an industry is a major determinant of a market's location along the market power spectrum. Markets with many firms (other things equal) are more competitive and are considered to have less concentrated industries. Markets with fewer firms (other things equal) are less competitive and are considered to have more concentrated industries.

In chapter 8, it was pointed out that imperfectly competitive market structures include **monopolistic competition** and **oligopoly** (see Figure 8.2). Monopolistically competitive markets have relatively low degrees of industry concentration whereas oligopolistic markets have relatively high degrees of industry concentration. Because the market power spectrum is continuous, we require a measure of industry concentration that allows us to identify the degrees of market power in different markets relative to one another. In this section, we will introduce two different measures of industry concentration that many economists use to compare different markets.

Before we construct these measures, we must introduce a very important definition that will serve as the building block for our measures of industry concentration. A firm's **market share** (S_i) refers to the percentage of the market's total sales revenue for which the i^{th} largest firm is responsible during a given period. For example, in a specific industry, S_2 refers to the second largest firm's market share. Suppose that the second largest firm's TR is \$10,000 per month (TR_2) and the total revenue earned in the entire market is \$100,000 per month (TR_M). In that case, S_2 is defined and calculated for this case as follows:

$$S_2 = \frac{TR_2}{TR_M} \cdot 100 = \frac{10{,}000}{100{,}000} \cdot 100 = 10$$

In this case, the second largest firm has 10% of the market. It should be noted that the final calculation is 10 rather than 10%. That is, the 10% is implied. This point is important because the correct calculations of the measures that we construct require the omission of the percentages to obtain the correct answer as we shall soon see.

The first measure of industry concentration that we will introduce is referred to as the **four-firm concentration ratio**. Simply put, it is the combined market share of an industry's four largest firms. It is defined as follows:

$$Four - Firm\ Concentration\ Ratio = S_1 + S_2 + S_3 + S_4$$

For example, suppose that an industry consists of the following market shares: 10, 20, 30, 25, and 15. Notice that these market shares sum to 100. Therefore, this industry contains exactly five firms. The four-firm concentration ratio in this case is 90 (= 30+25+20+15) because the smallest firm is excluded. This market appears to be highly concentrated because the four largest firms control 90% of the market.

The four-firm ratio appears to provide a sound measure of industry concentration. The larger it is, the more concentrated the industry is. The smaller it is, the less concentrated the industry is. A problem arises, however, when we consider examples of a specific sort. For example, consider two industries in which the market shares are 50 and 50 in Industry A and 25, 25, 25, and 25 in Industry B. In each case, the four-firm concentration ratio is 100. This result suggests that the two industries are equally concentrated. Our intuition, however, tells us that the industry with only two firms is significantly more concentrated than the industry with four firms. The four-firm ratio appears to be a deficient measure in the sense that it does not make distinctions between different industries that are as precise as we would prefer.

For this reason, we now introduce a second measure of industry concentration that is free from the defects of the four-firm ratio. The second measure of industry concentration is called the **Herfindahl-Hirschman Index (HHI)**. It is named after the American economist Orris Herfindahl and the German economist Albert Hirschman. This measure is defined as follows for n firms in an industry:

$$HHI = S_1^2 + S_2^2 + S_3^2 + ... + S_n^2$$

The reader should notice two key differences between the four-firm ratio and the HHI. Unlike the four-firm ratio, the HHI includes *every* market share in the industry, not just the market shares of the largest four firms. Additionally, each market share is squared before the terms are added together. To understand why these changes have been made, let's return to our example involving industries A and B from earlier. The HHI in Industy A (HHI_A) and in Industry B (HHI_B) are calculated as follows:

$$HHI_A = 50^2 + 50^2 = 5000$$
$$HHI_B = 25^2 + 25^2 + 25^2 + 25^2 = 2500$$

Whereas the four-firm ratio suggests that the two industries are equally concentrated, the HHI indicates that Industry A is twice as concentrated as Industry B. This result is much more consistent with our intuition and is an important reason to rely on the HHI, particularly when one is trying to make careful distinctions between levels of industry concentration. It should now be clear as well why the HHI includes the squaring of the market shares. The squaring technique allows the larger market shares to count much more heavily in the overall index than the smaller market shares. The inclusion of all market shares also allows us to obtain a more precise measure.

The HHI may also take on a much larger range of values than the four-firm ratio. That is:

$$10,000 \geq HHI > 0$$

In the extreme case in which a single firm has a market share of 100, the HHI equals 10,000 (= 100^2). This case is the case of pure monopoly, and it should be noted that this situation is the only one that leads to a HHI of 10,000. The four-firm ratio, however, is 100 just like we found for Industries A and B. The lower extreme for the HHI is just above zero. Clearly, the only way to have a HHI of 0 is for all firms to have zero market shares, which means that the market simply does not exist. Therefore, for any existing markets, the smallest HHI is just above zero. Such a market would be very close to being perfectly competitive due to the huge number of very small firms. It is important not to focus on the specific value of the HHI. The index value by itself means rather little. What is important is the index value of one industry relative to another. The higher the index value is, the more concentrated is the industry.

The Antitrust Division of the United States Justice Department considers a HHI below 1,500 to be an **unconcentrated** industry, a HHI between 1,500 and 2,500 to be a **moderate degree of concentration**, and a HHI above 2,500 to be a **high degree of concentration**.[1]

Using data from the Economic Census provided by the United States Census Bureau,[2] we consider a few examples of unconcentrated industries shown below:

- In 2002, the four-firm concentration ratio in animal food manufacturing was 39.3 and the HHI for the 50 largest firms was 636.6.[3]
- The 2002 four-firm ratio in flour milling and malt manufacturing was 38.2 and the HHI for the 50 largest firms was 524.4.
- The 2002 four-firm ratio in sugar manufacturing was 52.8 and the HHI for the 50 largest firms was 855.5.

A few examples of moderately concentrated industries are below:

- In 2002, the four-firm concentration ratio in beet sugar manufacturing was 85.3 and the HHI for the 20 largest firms was 2,208.9.
- The 2002 four-firm ratio in cookie and cracker manufacturing was 70.9 and the HHI for the 50 largest firms was 1,901.2.
- The 2002 four-firm ratio in tortilla

manufacturing was 59.3 and the HHI for the 50 largest firms was 2,355.4.

A few examples of highly concentrated industries are below:

- In 2002, the four-firm concentration ratio in tobacco stemming and redrying was 84.2 and the HHI for the 20 largest firms was 2,905.9.
- The 2002 four-firm ratio in house slipper manufacturing was 94.3 and the HHI for the 20 largest firms was 2,943.5.
- The 2002 four-firm ratio in glass container manufacturing was 87.1 and the HHI for the 50 largest firms was 2,548.1.

Notice that tobacco stemming and redrying appears less concentrated than glass container manufacturing when we compare the four-firm ratios for the two industries, but tobacco stemming and redrying appears more concentrated than glass container manufacturing when we consider the HHI for the two industries. The two measures sometimes lead to contradictory results, but as we have seen, the HHI leads to more precise results than the four-firm ratio. Nevertheless, the four-firm ratio has a greater intuitive meaning and so it is often considered as well when investigating the degree of concentration in an industry. Now that we have a method of measuring industry concentration, we can explore the characteristics of the imperfectly competitive market structures that fall in between the extremes of perfect competition and pure monopoly.

Characteristics of Monopolistic Competition

Monopolistically competitive markets have three major features, two of which suggest a competitive market structure and one of which suggests a more monopolistic market structure. The two characteristics that contribute to the competitive nature of this market structure are the relatively large number of sellers and the absence of barriers to entry and exit in the long run. The characteristic that contributes to the monopolistic nature of this market structure is the tendency of all firms in the industry to produce products with slight differences that they promote aggressively through advertising. It is important to recognize that the differences need not be real or actual differences. If the consumer perceives a difference to exist across products in an industry, we may regard the market structure as monopolistically competitive. For example, doctors may tell patients that no significant differences exist across different types of aspirin, but the companies that produce aspirin promote their specific brands with great vigor. If consumers believe the differences exist, then each firm will possess some degree of market power. Other examples of monopolistically competitive markets include the markets for restaurant meals and auto repairs. Sometimes the factor that differentiates one product from another is nothing more than the location of the seller.

One good example of a monopolistically competitive market is the market for celebrity scents. As of 2009, firms were releasing at least 500 different kinds of celebrity perfume and cologne each year with department stores selling millions of bottles. Slight product differentiation is the source of each firm's market power in this market. Firms have developed celebrity perfumes for Jennifer Lopez, Gwen Stefani, and

Sarah Jessica Parker. Rapper 50 Cent also has a fragrance called Power by 50. Fans that waited at Macy's in midtown Manhattan to buy a bottle of the new cologne had the good luck to have their photograph taken with the famous rapper. One fan exclaimed, "It's him, it's not the perfume. I can't explain it, it's like an energy you carry."[4] Clearly, the uniqueness of the product often stems as much from the perceptions of the consumer as it does from the characteristics of the product itself.

Because each firm in a monopolistically competitive market has some market power, each firm faces a downward sloping demand curve for its specific product as shown in Figure 10.1.

Figure 10.1: The Demand Curves Facing Two Different Firms

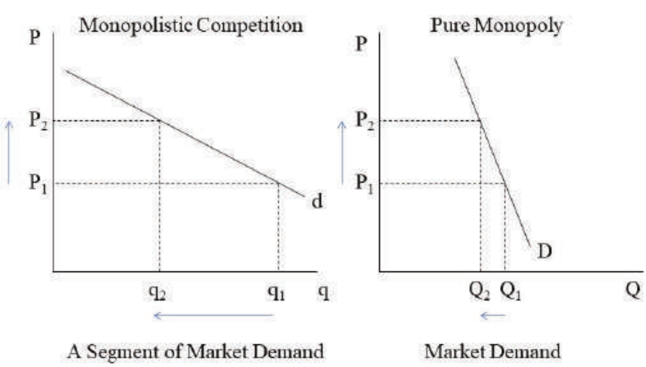


Figure 10.1 also shows the graph of a downward sloping demand curve facing a pure monopolist. Two important differences between the two graphs should be noted. First, whereas the demand curve facing the pure

monopolist is the entire market demand curve, the demand curve facing the monopolistically competitive firm is only a segment of the market demand. Second, the pure monopolist tends to face a much more inelastic demand curve because no close substitutes for the product exist. Therefore, consumers will be much less responsive to price changes when the market is purely monopolistic because they cannot find good substitute products. A monopolistically competitive firm faces competition from firms that produce close substitutes. The widespread availability of substitutes suggests that the demand curve facing the monopolistically competitive firm will be far more elastic since consumers can respond to price changes more easily.

Aside from these differences, the short run analysis of profit maximization for a monopolistically competitive firm is very similar to that of a purely monopolistic firm. Figure 10.2 shows two possible short run scenarios for the monopolistically competitive firm.

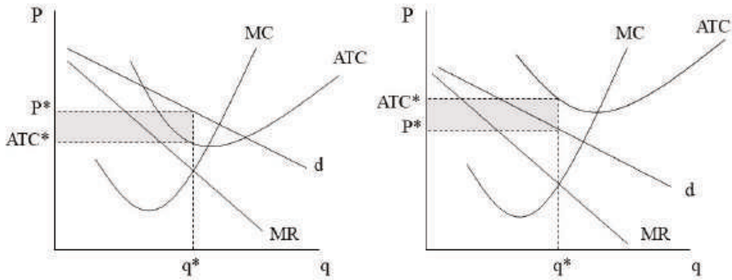


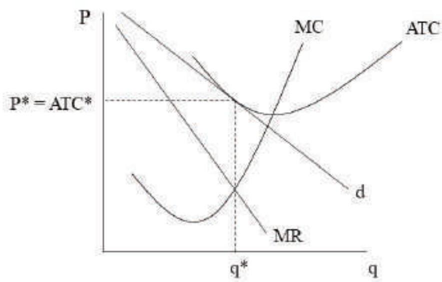
Figure 10.2: Short-Run Equilibrium Positions for a Monopolistically Competitive Firm

Positive Economic Profits (P > ATC)   Negative Economic Profits (P < ATC)

Depending on the relationship between the demand curve facing the firm and the unit cost curves, the firm may experience positive economic profits or negative economic profits in the short run. If the demand facing the firm is relatively high and unit costs are relatively low, then the firm will enjoy short run economic profits. If the demand facing the firm is relatively low and unit costs are relatively high, then the firm will suffer short run economic losses.

Long run profit maximization for the monopolistically competitive firm is a very different story. Because firms are free to enter and exit the market in the long run, short run economic profits or losses will be eliminated in the long run. The long run result is, therefore, an economic profit of zero. This situation is depicted in Figure 10.3.

Figure 10.3: A Zero-Economic Profit Long-Run Equilibrium for a Monopolistically Competitive Firm (P = ATC)



The reader should notice in Figure 10.3 that the demand curve facing the firm is just tangent to the ATC curve above the MR = MC intersection. At this point, price and ATC are the same and so economic profits are equal to zero in this case. How does this long run adjustment to zero economic profits occur? Two possible scenarios exist:

1. Economic profits exist in the short run, leading to the following:

- Firms enter the industry in the long run.
- The firms in the industry lose market share to the entering firms causing the individual demand curves that they face to shift leftward.
- This adjustment continues until P = ATC and economic profits are equal to zero.

2. Economic losses exist in the short run, leading to the following:

- Firms exit the industry in the long run.
- The remaining firms in the industry gain market share causing the individual demand curves that they face to shift rightward.
- This adjustment continues until P = ATC and economic profits are equal to zero.

The consequences of monopolistic competition are relatively straightforward. First, monopolistically competitive markets lead to inefficiency. Markup pricing exists (P > MC), which monopolistically competitive firms achieve by restricting their production levels below those that would prevail in a perfectly competitive market. Furthermore, least-cost production is not achieved because ATC exceeds the minimum ATC. In Figure 10.3, ATC* is higher than the minimum ATC that exists at the intersection between MC and ATC. The monopolistically competitive firm thus operates at less than optimal capacity. The conclusion that monopolistically competitive firms have a chronic problem of under-utilization of plant capacity is referred to as the **excess capacity theorem of monopolistic competition**.

Recall that the more unique the product is (i.e., the more differentiated it is), the steeper the demand curve will be because fewer close substitutes exist for the product. Similarly, if less product differentiation exists, then the demand curve is flatter because more close substitutes exist. In the extreme case that zero product differentiation exists, the market is once again perfectly competitive and the demand curve is perfectly horizontal. In that case, the MR = MC intersection will occur at the minimum point on the ATC curve. Hence, the problem of inefficiency in a monopolistically competitive market worsens as the products become increasingly differentiated. The increasing amount of differentiation gives individual firms more market power, which they use to restrict output further to raise prices and economic profits.

On the other hand, most consumers value some product differentiation. For example, if all firms in the shoe industry only produce one style of shoe, then the market will be perfectly competitive. A larger number of shoes would be produced at a lower per unit cost and price. Would our society suffer a different kind of loss, however, due to everyone wearing the same style of shoe? It would seem to be so. We would lose the ability to express our uniqueness in this way. Most of us would be willing to pay something extra for that opportunity of self-expression. How much more we are willing to pay for product variety is the question that is highly debatable. Our answers will surely vary depending on the specific good in question. A tradeoff thus exists between product variety and economic efficiency.

Characteristics of Oligopoly

Oligopolistic markets have several defining characteristics. These are markets dominated by a few large firms. Oligopolistic markets may have just two or three firms or even nine or ten firms. The specific number of firms is not rigidly defined. Again, the degree of concentration in an industry is measured along a continuum. For decades, the U.S. automobile market was dominated by the “Big Three” (i.e., General Motors, Ford, and Chrysler). In the beer industry, the “Big Five” in the early 1990s included Anheuser Busch, Miller, Stroh, Coors, and Heilemann. Oligopolies have also existed in the markets for breakfast cereal, steel, and military weaponry.

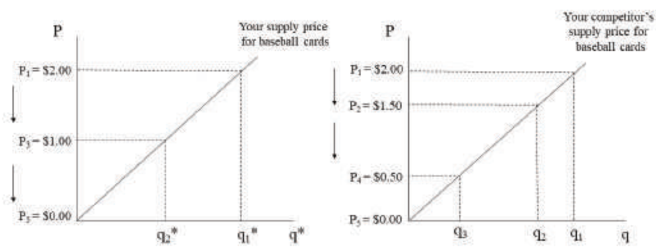
Also required for a market to be oligopolistic is that each firm has considerable market power, which is the result of barriers to entry that discourage competition. Such entry barriers might include patents, or large startup costs that generate economies of scale. The final characteristic of oligopoly markets is that the firms in the industry behave strategically when setting prices and output levels. That is, each firm sets its price while considering the prices and likely reactions of its competitors. This characteristic sharply distinguishes this market structure from the other market structures. In a perfectly competitive market structure, each firm takes the price as given and ignores the actions of competitors. In a purely monopolistic market, no competitors exist. In a monopolistically competitive market, firms might strive to differentiate their products, but at a point in time, the prices they charge are entirely determined by the demand segments they face and the production costs they incur.

The element of strategic interaction makes it very difficult to model oligopoly behavior. As a result, no single oligopoly model is dominant within neoclassical economics. We will consider several models as well as explore how a special subfield of economics called **game theory** may shed light on the behavior of oligopolistic firms.

A Simple Duopoly Model

Sometimes oligopoly markets only contain two firms. These special cases of oligopoly markets are called **duopoly markets**. Suppose that you and your competitor are the only two sellers in the market for baseball cards. Your supply curve and your competitor's supply curve are shown in Figure 10.4.[5]

Figure 10.4: A Price War Erupts in a Duopoly Market



The **supply price** refers to the minimum price that these sellers are willing and able to charge for the corresponding output level. Initially, your supply price and your competitor's supply price are P_1 = \$2.00 per baseball card and your output levels are each q_1. Now assume that you and your competitor think strategically about each other's price. Let's assume that pricing takes place in a series of rounds with each seller reacting to the other seller's price every other round. Further suppose that you and your competitor each decide that you will always cut the price per card by \$0.50 below the price set by the competition. Finally, assume that your competitor is the first to act. Your competitor cuts the price to P_2

= \$1.50, undercutting you by \$0.50. Since you follow a similar strategy, you decide to cut your price to P_3 = \$1.00, thus cutting your price in half and undercutting your competitor by \$0.50. In response, your competitor cuts price to undercut your new price and so sets a price of P_4 = \$0.50. Your stubborn adherence to this cutthroat strategy leads you to reduce your price to zero and your competitor follows. If the two sellers adhere strictly to these pricing rules, a zero price/zero quantity equilibrium is the result since quantities supplied decline with the price cuts.

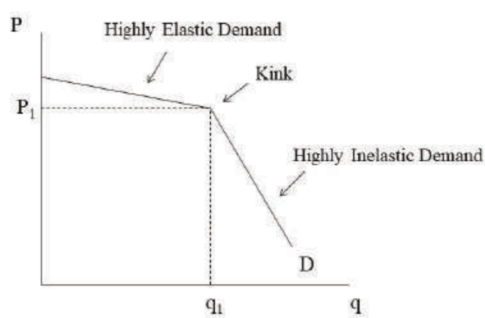
The problem with this model, of course, is that actual oligopoly markets do manage to persist without price wars completely undermining production. For one thing, firms have production costs and so they may not be willing to cut prices so low that they suffer extensive losses. Also, firms are able to grab a larger market share when they undercut competitors and that factor is not taken into account in Figure 10.4. Therefore, we seem to require models that can explain how oligopoly firms may coexist while accounting for the actions of their competitors and earning positive economic profits. It is to these models that we now turn.

The Kinked Demand Model of Non-Collusive Oligopoly

In 1939, the Marxian economist Paul Sweezy introduced what has become known as the **kinked demand curve model** of non-collusive oligopoly.[6] It is rare that a Marxian economist develops a model that becomes widely taught by neoclassical economists. The kinked demand curve model is an important example of how sometimes ideas developed within one school of economic thought spill over into other schools of economic thought. Certainly, not all neoclassical economists accept the validity of this model's conclusions, but it has been incorporated into most mainstream economics textbooks.

The purpose of the model is to explain how oligopoly firms manage to establish stable prices in the marketplace without relying on collusion. That is, price wars are somehow avoided in these markets even though the firms never coordinate with one another to explicitly fix prices. Because the purpose is to explain the rigidity or "stickiness" of the price rather than the level of the price itself, the model assumes that one of the oligopolistic firms is already charging a specific price P_1 and producing a specific quantity q_1 as shown in Figure 10.5.

Figure 10.5: The Kinked-Demand Curve Facing One Firm

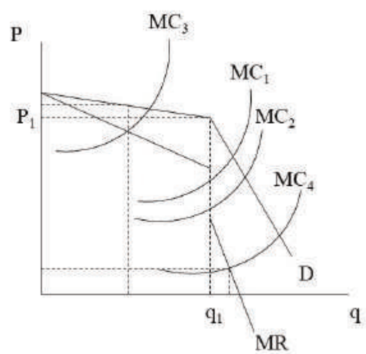


Given the original price of P_1, we need to determine the shape of the demand curve facing this firm. To derive the demand curve facing the firm, we ask what will happen if this firm raises its price above P_1. Because we are assuming that firms think strategically, it is reasonable to expect that when this firm raises its price, the other firms will *not* follow with a price increase. By not following with a price increase, they can capture a larger share of the market. As a result, when this firm raises its price, its customers will react by purchasing the product from the firm's competitors. That is, consumers will be very responsive to the price increase and the quantity demanded will fall significantly. In other words, the demand curve facing the firm is very elastic above the price of P_1.

On the other hand, suppose this firm reduces its price below P_1. Because we are assuming that firms think strategically, it is reasonable to expect that when this firm cuts its price, the other firms *will* follow with a price cut. By following with a price cut, the other firms can prevent this firm from capturing a larger share of the market at their expense. As a result, when this firm cuts its price, it will not be able to expand its sales much at all. That is, consumers will be very unresponsive to the price cut and the quantity demanded will only increase a small amount due to new customers entering the market. In other words, the demand curve facing the firm is very inelastic below the price of P_1.

Because the demand curve is relatively flat above P_1 and relatively steep below P_1, the demand curve facing the firm has a kink at a price of P_1. The next step is to determine the firm's marginal revenue (MR) curve. We have seen that a downward sloping demand curve has a corresponding MR curve that declines more quickly than demand. Again, the reason is that when the firm cuts the price to sell another unit, it must cut the price of all units sold. As a result, the revenue rises by less than the price of the additional unit. In the case of a kinked demand curve, we essentially observe two downward sloping demand curves joined at the kink. Each section then should have a downward sloping MR curve that corresponds to it. Figure 10.6 shows the MR curve for an oligopolistic firm facing a kinked demand curve.

Figure 10.6: The Marginal Revenue Curve of a Firm Facing a Kinked Demand Curve



In Figure 10.6, the MR curve has a gap in it. The gap occurs at the same output level of q_1 at which the kink occurs. This gap in the MR curve can be explained intuitively. Imagine that the firm is cutting the price at levels above P_1. The MR is given by the flatter portion of the MR curve. Once the firm reaches P_1, a substantial price cut is needed to sell one more unit. This large price cut causes the price to drop considerably for all other units sold and so the MR plummets, creating a gap in the MR curve. If the firm then continues to cut price along the inelastic portion of the demand curve, the MR falls steadily along the steeper portion of the MR curve.

Several marginal cost (MC) curves have also been included in Figure 10.6. The MC curve of the oligopolistic firm is determined by the very same factor discussed previously: production technology. That is, the MC declines due to specialization and division of labor and then rises due to diminishing returns to labor. We can now determine the profit-maximizing output and price. Following the rule that the firm will equate MR and MC, if the firm's MC is MC_1 then the firm will produce where MR and MC_1 intersect. Technically, a discontinuity occurs in the MR curve and so no intersection occurs. Despite this technicality, the firm will increase its profit by producing more when MR exceeds MC, and the firm will increase its profit by producing less when MC exceeds MR. This tendency leads the firm to produce at an output level of q_1. The price charged is then set up on the demand curve above this output level at P_1. Interestingly, we obtain the same result if marginal cost declines to MC_2. That is, the firm will set price and output at the kink in the demand curve. The implication is that unit cost can fluctuate significantly with no effect on price or output. That is, the model provides an explanation for price rigidity in oligopolistic markets. Even if unit cost drops significantly, firms will hesitate to cut price because doing so will cause a price war. Similarly, even if unit cost rises significantly, firms will hesitate to raise price because doing so will cause a loss of market share. Price changes, therefore, lead to a reduction in the firm's economic profits. If all firms reason in a similar manner, then price stability arises even without collusion.

It should be noted that price changes are possible in the kinked demand model. A rise in price might occur, for example, if MC rises a great deal. A rise in MC to MC_3 will lead to a reduction in output below q_1 and a rise in price above P_1. A drop in MC to MC_4 will lead to an increase in output above q_1 and a decrease in price below P_1. The point is not that prices never change in oligopolistic markets but only that they are less likely to change than in other market structures when production costs change. If you think about Coke and Pepsi, for example, these two firms do not collude. If they fixed prices, it would be illegal. Still, relatively stable and similar prices emerge in the market for soda. The author encountered another interesting case a few years ago at Wendy's. A tomato shortage had caused the price of tomatoes to rise significantly. As a result, Wendy's asked each customer who ordered a hamburger whether they would be willing to forego the slice of tomato that typically is included on a Wendy's hamburger. In this case, the firm's MC rose, and it made a special effort to enlist the help of customers in keeping MC down, all to avoid a price increase that might cost it market share if its competitors refused to follow the price increase.

The kinked demand curve model of oligopoly is useful for thinking about how stable prices may emerge in oligopolistic markets. It is subject to the criticism, however, that it only explains the stability of the product price and not its magnitude. For that reason, neoclassical economists have several other ways of thinking about oligopolistic behavior.

Two Additional Models: Collusive Oligopoly and Price Leadership

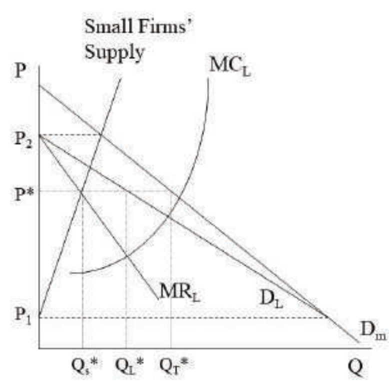
The kinked demand model of non-collusive oligopoly assumes that the firms do not explicitly agree on price. Each firm chooses its price and output independently even though each carefully considers what its competitors are doing and will do in response to a price change. If firms do collude, then essentially the firms will act as a pure monopoly. Such a collusive oligopoly is referred to as a **cartel**. Quite simply, a cartel is a group of firms that behaves as a single firm with respect to certain decisions, which might include the product price, the output level, or the market share for each cartel member. The most famous example of a cartel is probably the Organization of Petroleum Exporting Countries or OPEC. This international oil cartel, made up primarily of Middle Eastern oil producers, has manipulated oil prices at different times since it was first formed in the 1960s.

The major difference between a cartel and a purely monopolistic firm is that a cartel is much less likely to remain intact over time for several reasons.[7] First, cartel members are often located in different geographic regions. As a result, the individual demands that they face and the production costs they incur may be significantly different. As a result, it will be difficult for the firms to agree on an appropriate product price. Second, cartel agreements tend to be weaker when a larger number of firms belong to the cartel. That is, it is easier to reach and maintain a pricing agreement among three firms than it is among eight firms. Third, the temptation to cheat on the pricing arrangement is immense because a firm can capture a greater market share by secretly undercutting the other cartel members. When all firms begin to cheat, the cartel falls apart. Such temptations are greatest during periods of economic recession when falling demand leads to desperate attempts to attract new customers. Fourth, cartels are often slow to form because the firms in the market fear that a high cartel price will attract new competition in the market. Finally, the antitrust laws discussed in the last chapter serve as a major barrier to cartel formation in the United States. Firms accused of price fixing face federal prosecution, steep fines, and even criminal punishment for their efforts to rig markets in their favor.

In other markets, market-sharing agreements may emerge in which one large firm acts as the price leader and many small competitors follow along by setting the same price as the leader. The price leader ends up sharing the market with many small firms by means of an implicit agreement. For example, U.S. Steel acted as the price leader in the steel industry during the early twentieth century. U.S. Steel would publish price catalogs periodically in which it would announce its new prices. Smaller steel companies would inevitably follow along. When the steel companies deviated from this policy during the panic of 1907, U.S. Steel retaliated with even steeper price cuts that severely punished the smaller steel companies. With this implicit threat in place, the steel companies quickly learned that they must continue to follow the lead of U.S. Steel in the pricing of steel products. In the U.S. automobile industry in the mid-twentieth century, a similar pattern arose with General Motors serving as the price leader. The president of GM would announce new automobile prices in a speech and the other automakers would fall into line with similar prices.

Neoclassical economists use a **price leadership model** to investigate how the established price, market demand, and the cost structures of the different firms collectively determine how the market is divided between the price leader and the smaller competitors. Figure 10.7 provides a graph that represents the price leadership model for (let's say) the automobile market.[8]

Figure 10.7: The Price Leadership Model



This graph is a bit involved and so we will discuss each piece before bringing all of them together to provide an explanation of how the market is divided when a price leader exists. First consider the market demand for automobiles (D_m). It is a downward sloping curve consistent with the law of demand. The small firms' supply curve, on the other hand, is upward sloping consistent with the law of supply. As the price increases, the quantity supplied of the small firms rises because the higher price covers the higher marginal cost of production. If no other firms exist in this market, then the market will be perfectly competitive and the equilibrium price will be the competitive price of P_2 where the supply of the small firms and market demand intersect.

Another firm does exist in this market, however, and it is the price leader. We wish to know what the demand curve is that faces the price leader in this market. We can derive the demand curve facing the price leader (D_L) by asking how much of the market demand remains at each price after the small firms have selected their quantity supplied. For example, at a price of P_2 the small firms produce enough to satisfy the entire market demand. Nothing remains then for the price leader and so the quantity demanded of the price leader's product is zero at a price of P_2. This point is one point on the demand curve facing the price leader. On the other hand, at a price of P_1, the small firms produce nothing. The quantity demanded facing the price leader in that case is the entire quantity demanded in the market. This point is a second point on the demand curve facing the price leader. As the price declines from P_2 to P_1, the small firms' quantity supplied falls continuously causing the price leader to acquire an increasing amount of the market demand until the price leader captures the entire market. The demand curve facing the price leader is, therefore, a downward sloping demand curve connecting the two points just mentioned.

Once we have determined the demand curve facing the price leader, it is a short step to obtain the MR curve for the price leader (MR_L). It is also downward sloping and falls more quickly than the demand curve for the very same reasons provided previously whenever we have derived the MR curve from a downward sloping demand curve. Additionally, the MC curve for the price leader (MC_L) is upward sloping due to the rising marginal cost of production, which stems from diminishing returns to labor.

We now have all the component parts, which we can bring together to complete the price leadership model. The price leader considers the situation and aims to maximize its economic profit. It sets MR_L equal to MC_L, as any profit-maximizing firm does, and produces Q_L*. To sell this amount of output, it must set the price up on the demand curve facing the price leader at P*. If the price leader sets the price up on the market demand curve then the price leader will not be able to sell Q_L* and the firm's economic profit will not be maximized. The small firms follow the price leader and set the same price of P*. The quantity supplied of the small firms shows up in two different ways in this graph. At a price of P*, we can look at the small firms' supply curve to see that the small firms will supply Q_S* at that price. Equivalently, the small firms produce enough to satisfy the market demand that remains after the price leader has set its output. Because Q_T* is the quantity demanded in the entire market at a price of P*, Q_S* must also equal the difference between Q_T* and Q_L*. That is, the small firms produce what is left over of the market demand after the price leader has taken its share of the market. We can express this result symbolically as follows:

$$Q_S^* = Q_T^* - Q_L^*$$

The market is thus divided between the price leader and the small firms. Clearly, the price leader could capture the entire market by setting the price all the way down at

P_1. The price leader has no incentive to do so, however, because it would produce far more output than is profit-maximizing. If the firm produced at that output level, MC would far exceed MR. Contrary to what one might expect, it is profit-maximizing for the price leader to share the market with the smaller firms.

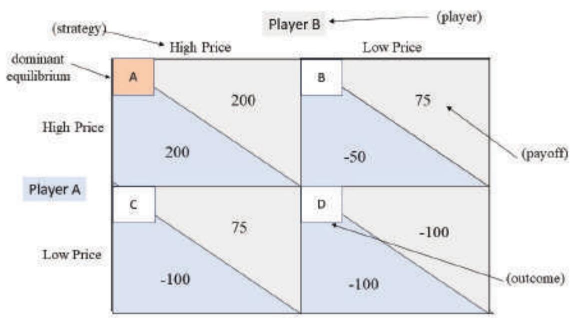
In general, oligopoly markets tend to be inefficient. In the kinked demand model, the cartel model (i.e., the pure monopoly model), and the price leadership model, the firms set price above MC. All such firms restrict output to maximize their economic profits. On the other hand, the large economic profits of oligopolistic firms enable them to invest in new production technologies that can enhance efficiency. Whether and to what extent these investments offset or cancel out the inefficiencies of oligopolistic firms is a question that may vary by industry.[9]

A Detour into Game Theory

Neoclassical economists also use an analytical method referred to as game theory to investigate the competitive interaction of oligopolistic firms. Game theory has been applied to many other topics in economics as well as other fields, including biology and political science. It is an analytical method that makes possible the study of strategic interaction between individual agents or **players**. In this section, we will discuss **simultaneous games**, in which all players simultaneously choose their **strategies** or courses of action.

The first game that we will consider has only two players: Player A and Player B. Each player has two possible strategies from which to choose. That is, a player may set a high price or a low price. Figure 10.8 shows a **payoff matrix** with four possible **outcomes**.

Figure 10.8: A Duopoly Game (Example #1)



If both players set a high price, then outcome A is the result. If both players set a low price, then outcome D is the result. If player A sets a high price and player B sets a low price, then outcome B is the result. Finally, if player A sets a low price and player B sets a high price, then outcome C is the result. Each outcome has a set of **payoffs** associated with it, which may be interpreted as the economic profits that each player receives if that outcome occurs. If the payoff is negative, then it may be interpreted as an economic loss.

Our purpose is to determine which of the four outcomes will occur when the players simultaneously choose their strategies. The reader may be tempted to select outcome A because both players receive the highest payoffs when this outcome occurs. Although in this case outcome A is the correct outcome, this approach frequently does not lead to the correct outcome. The following approach may be used to arrive at the correct outcome:

1. If Player A sets a high price, then Player B sets a high price (since 200 > 75).
2. If Player A sets a low price, then Player B sets a high price (since 75 > -100).

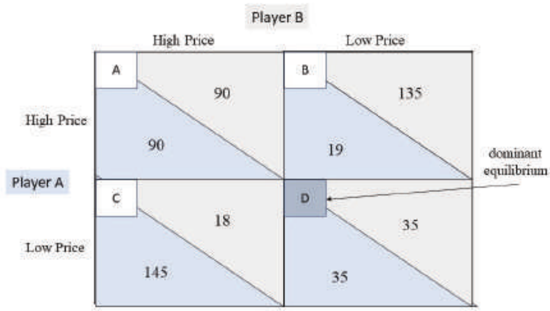
In this case, we say that a high price is a **dominant strategy** for Player B, meaning that the **best response** of Player B is always to select the same strategy regardless of what Player A chooses to do. We now investigate Player A's best responses to Player B's possible strategies.

1. If Player B sets a high price, then Player A sets a high price (since 200 > -100).
2. If Player B sets a low price, then Player A sets a high price (since -50 > -100).

In this case, Player A also has a dominant strategy, which is to always set a high price. Both players, therefore, have dominant strategies. Because both players will choose to set a high price, the outcome of the game is outcome A, and each player receives a payoff of 200. When the equilibrium outcome of the game results from both players having dominant strategies, the outcome is called a **dominant equilibrium**. Outcome A is, therefore, a dominant equilibrium.

Consider the second game represented in Figure 10.9.

Figure 10.9: A Duopoly Game (Example #2)



We will follow the same procedure that we used with the first game to determine the outcome of the game.

1. If Player A sets a high price, then Player B sets a low price (since 135 > 90).
2. If Player A sets a low price, then Player B sets a low price (since 35 > 18).

In this case, Player B has a dominant strategy to always set a low price regardless of the strategy that Player A chooses. Furthermore:

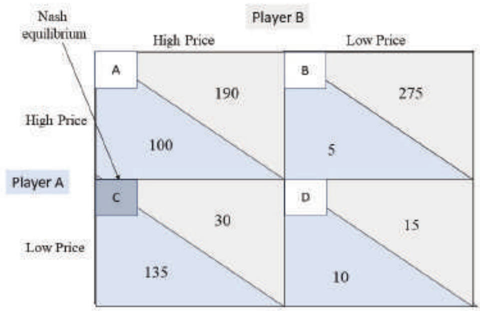
1. If Player B sets a high price, then Player A sets a low price (since 145 > 90).
2. If Player B sets a low price, then Player A sets a low price (since 35 > 19).

In this case, Player A also has a dominant strategy to always set a low price regardless of the strategy that Player B chooses. Since both players opt to set a low price, the outcome of the game is outcome D, and each player receives a payoff of 35. Because both players have dominant strategies, the outcome of the game is a dominant equilibrium.

The outcome of this game is especially interesting because each player receives a lower payoff than he or she would have received had both players set high prices. Essentially, what happens here is that a price war leads to a **suboptimal outcome** for both players. The reader might conclude that collusion would prevent this problem from arising. If each player agrees to charge a high price, then outcome A can be achieved. Is it true? Suppose the two players agree to charge a high price. Each player knows that its best response to the other player charging a high price is to charge a low price. Player A, for example, will enjoy a payoff of 145 rather than 90 by undercutting Player B. Similarly, Player B will enjoy a payoff of 135 rather than 90 when undercutting Player A. In other words, both players have an incentive to cheat on their pricing agreement. Since both players aim to maximize their payoffs, the pricing agreement falls apart. Game theory thus sheds light on the reason that cartel agreements are difficult to maintain except for short periods. The incentive to cheat is simply too great.

Consider a third game represented in Figure 10.10.

Figure 10.10: A Duopoly Game (Example #3)



Again, we will follow our method of determining the outcome of the game.

1. If Player A sets a high price, then Player B sets a low price (since 275 > 190).
2. If Player A sets a low price, then Player B sets a high price (since 30 > 15).

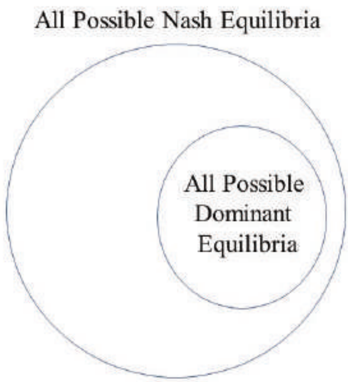
We see here that Player B does not have a dominant strategy since Player B's best response depends upon Player's A's choice of strategy. Furthermore:

1. If Player B sets a high price, then Player A sets a low price (since 135 > 100).
2. If Player B sets a low price, then Player A sets a low price (since 10 > 5).

Player A clearly has a dominant strategy to always set a low price regardless of which strategy Player B chooses. Still, in this case, the outcome is not immediately obvious since Player B does not have a dominant strategy. Nevertheless, we can arrive at the outcome if we assume that each player is able to anticipate the decision-making process of the other player. For example, if Player B can see that Player A will always set a low price no matter what Player B chooses, then Player B will choose her best response to Player A setting a low price. In this case, Player B will set a high price because setting a high price is Player B's best response to Player A setting a low price. Outcome C is the result of this game.

Because an outcome is called a dominant equilibrium only when both players have dominant strategies, outcome C is not a dominant equilibrium. Nevertheless, the outcome is a **Nash equilibrium**, which is a concept named after John Nash for his discovery of it in the 1950s and for which he was awarded the Nobel Prize in Economics in 1994. A Nash equilibrium exists when every player has chosen her best response given the strategies of the other players. Outcome C (A low and B High) is a Nash equilibrium because Player A's best response is to always set a low price. Similarly, Player B's best response to Player A's low price is a high price. Clearly, not every Nash equilibrium is a dominant equilibrium as this example shows, but is every dominant equilibrium a Nash equilibrium? The answer is yes. A dominant equilibrium exists whenever both players have selected their dominant strategies and these strategies are always their best responses to whatever the other players have chosen. Figure 10.11 shows the relationship between the set of all possible Nash equilibria and the set of all possible dominant equilibria.

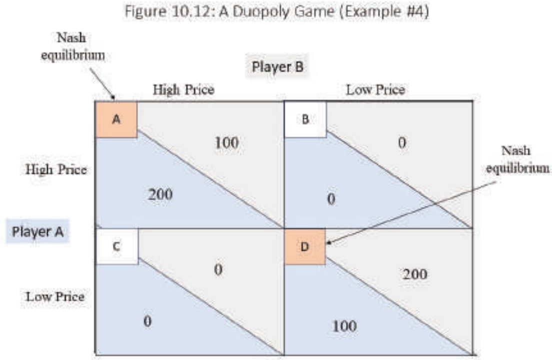
Figure 10.11: The Sets of All Possible Nash Equilibria and Dominant Equilibria



As the Venn diagram shows, every dominant equilibrium is a Nash equilibrium but not every Nash equilibrium is a dominant equilibrium.

The next game that we will consider has a surprising outcome. Consider this fourth game represented in Figure 10.12.

Figure 10.12: A Duopoly Game (Example #4)



Again, we will follow our method to determine the outcome of the game.

1. If Player A sets a high price, then Player B sets a high price (since 100 > 0).
2. If Player A sets a low price, then Player B sets a low price (since 200 > 0).

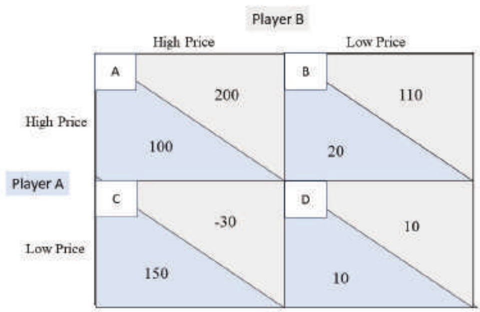
In this case, Player B does not have a dominant strategy. Furthermore:

1. If Player B sets a high price, then Player A sets a high price (since 200 > 0).
2. If Player B sets a low price, then Player A sets a low price (since 100 > 0).

Player A also lacks a dominant strategy. It will not be possible to determine the outcome of the game as we did in the third game because neither player can be sure what the other player will do. Nevertheless, it is possible to identify two Nash equilibria in this game. Notice that if Player A sets a high price, then Player B's best response is to set a high price. Similarly, if Player B sets a high price, then Player A's best response is to also set a high price. Therefore, when each player sets a high price, each player has chosen her best response to the other player. Outcome A is thus a Nash equilibrium. It should also be noticed that if Player A sets a low price, then Player B's best response is to set a low price. Similarly, if Player B sets a low price, then Player A's best response is to set a low price. When each player sets a low price then, each player has chosen her best response. Outcome D is thus also a Nash equilibrium. This game, therefore, has two Nash equilibria. Outcomes B and C are not Nash equilibria because neither player chooses her best response for those outcomes. Games with multiple Nash equilibria are thus possible.

The final game we will consider also has a surprising outcome. Consider the fifth game represented in Figure 10.13.

Figure 10.13: A Duopoly Game (Example #5)



Again, we will follow our method to determine the outcome of the game.

1. If Player A sets a high price, then Player B sets a high price (since 200 > 110).
2. If Player A sets a low price, then Player B sets a low price (since 10 > -30).

In this case, Player B does not have a dominant strategy since her best response depends upon Player A's choice of strategy. Furthermore:

1. If Player B sets a high price, then Player A sets a low price (since 150 > 100).
2. If Player B sets a low price, then Player A sets a high price (since 20 > 10).

Again, neither player has a dominant strategy. Therefore, the outcome cannot be determined as in the third game. Can we identify any Nash equilibria though? Let's see. If Player A sets a high price, then Player B will set a high price, but if Player B sets a high price, then Player A sets a low price. When considering this analysis of Figure 10.13, we fail to find the agreement we found when looking at the fourth game. Similarly, if Player A sets a low price, then Player B sets a low price, but if Player B sets a low price, then Player A sets a high price. Again, considering this analysis of Figure 10.13, we fail to find the agreement we found in the fourth game. Hence, no Nash equilibrium exists in this game. An alternative method of verifying that no Nash equilibrium exists is to consider each of the four outcomes in the game. For each outcome, we should ask whether each player has chosen her best response to the other player. If we answer in the negative for either player, then that outcome cannot be a Nash equilibrium. As the reader can see, these games have many possible outcomes!

A Post-Keynesian Markup Pricing Model

Post-Keynesian economists have a different model of oligopolistic behavior than the neoclassical models that we have explored in this chapter. As explained in Chapter 1, post-Keynesian economics is a complex body of thought that derives mostly from the ideas of John Maynard Keynes but which also contains elements of classical political economy, Marxian economics, Sraffian economics, and neoclassical economics. In this chapter, we will explore the post-Keynesian theory of pricing in oligopolistic markets in such a way that we see how some of these elements are brought together to provide a unique explanation of firm behavior.

Post-Keynesian economists distinguish between two types of markets that they refer to as “**flexprice**” markets and “**fixprice**” markets.[10] The flexprice markets are intensely competitive and prices are determined according to the laws of supply and demand.[11] The fixprice markets, on the other hand, are oligopolistic and have prices determined as the sum of the “normal cost of production” and a **markup** for profit.[12] Whereas flexprice markets tend to include markets for agricultural commodities and raw materials, fixprice markets include markets for finished, manufactured commodities.[13] For post-Keynesian theorists, the latter type of market is the most important in terms of understanding modern capitalist economies.[14]

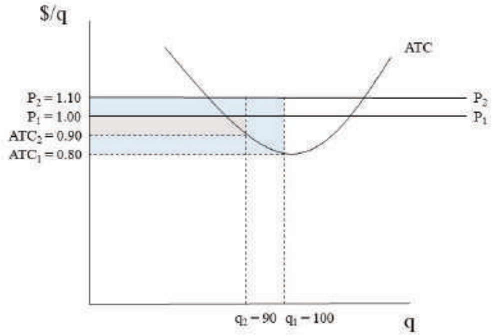
The key issue in fixprice markets is the determination of the profit markup.[15] Post-Keynesian economists argue that the profit markup is established according to firms’ need for internal financing of new investment projects out of current profits.[16] Their aim is to maximize sales revenue or market share rather than short run profit.[17] That is, if firms anticipate greater future demand for their products, then they will increase the profit markup. The larger markup will generate larger profits which can be used to expand production plant capacity. Similarly, firms will reduce the profit markup if they anticipate a reduction in future demand for their products. The reduction will allow them to reduce the level of investment and avoid creating unnecessary additional plant capacity.[18] The change in the markup will cause the price to change in these markets. The only other factor that can cause the price to change in these markets is a change in the normal level of production cost.[19] Why then are these markets referred to as *fixprice* markets? Because *short run* changes in *current* demand do not affect the product price. Instead, firms will expand production (or output) to meet short run surges in demand and will reduce output in response to short run drops in demand. Firms are assumed to possess some excess production capacity that allows them to make these short run adjustments. In flexprice markets, on the other hand, short run changes in demand do affect product prices, as explained in Chapter 3 in the context of the neoclassical supply and demand model.

It is possible to represent the post-Keynesian theory of oligopolistic pricing more precisely. Let's assume that the price is determined as the sum of short run unit cost (ATC) and the per unit profit markup as shown below:

$$P = ATC + markup$$

In this case, the ATC may be interpreted as the normal cost of production in the short run.[20] Figure 10.14 depicts the post-Keynesian markup pricing model.

Figure 10.14: A Post-Keynesian Markup Pricing Model



Suppose the firm initially charges a price of $1.00 per unit and produces an output of 100 units. The ATC is $0.80 per unit in this case. The markup is $0.20 per unit. Notice that the firm has a small degree of excess capacity because it has not quite reached the lowest point on its ATC curve. The firm's total profit in this case is $20 (= $0.20 per unit times 100 units). If a short run reduction in demand occurs, then the firm will maintain the price of $1.00 per unit but will reduce output from 100 units to 90 units. The firm keeps the price the same because this market is a fixprice market and short run changes in demand do not affect price. Post-Keynesian economists explain the reluctance to change price as stemming from firms' fear that they might "spoil the market by damaging customer goodwill." [21] The firm experiences an increase in excess plant capacity because it has reduced its output. Due to the failure to fully utilize its plant capacity and the gains from specialization, the per unit cost increases to $0.90 per unit. The markup thus falls to $0.10 per unit and profits decline to $9.00 (= $0.10 per unit times 90 units).

Suppose next that the firm expects an increase in the future demand for its product. The managers believe that plant capacity must be increased to meet the larger future demand. As a result, the firm increases its profit markup from $0.10 per unit to $0.20 per unit. This change will allow the firm to earn enough additional profit to finance the plant expansion using internal funds. The firm thus increases its price to $1.10. The firm's profits now increase to $18 (= $0.20 per unit times 90 units). If a short run increase in demand restores current demand to its previous level, now the firm's profits increase to $30 (= $0.30 per unit times 100 units). The firm is now charging a price of $1.10 per unit and incurring unit costs of only $0.80 per unit, generating a per unit profit of $0.30 per unit on these 100 units.

This analysis shows how oligopolistic firms may alter their production levels in the short run while maintaining constant prices in the face of changing current demand. At the same time, it shows how prices may be changed to support investment projects that will help satisfy higher expected future demand. The analysis also shows how profits vary at a microeconomic level as current demand rises and falls. This pattern of rising and falling profits with changing demand conditions is perfectly consistent with the post-Keynesian analysis of the business cycle that we will discuss in Chapter 14. The consistency of post-Keynesian microeconomic analysis and macroeconomic analysis is a point in its favor given how much difficulty neoclassical economists have faced with their efforts to create a synthesis of microeconomic and macroeconomic theory.

The model also shows how post-Keynesian economics is a blend of different bodies of thought. The belief that markets often fail to clear automatically is a central feature of Keynes's perspective. The suggestion that prices are determined primarily by production cost is consistent with classical economics. The assumption that labor specialization influences per unit cost is a feature of neoclassical cost analysis. Finally, the assumption that different rates of profit in different industries (stemming from long run changes in demand) lead to capital expansions in high-profit industries and capital contractions in low-profit industries is certainly a feature of Marxian economics. We will learn more aspects of post-Keynesian theory when we turn to macroeconomics.

Following the Economic News

As Kimberly-Clark, the maker of Huggies diapers, and Proctor & Gamble (P&G), the maker of Pampers, increased their competition in 2002, Kimberly-Clark tried to raise the effective price of its product by reducing the number of diapers in its packages, and cutting its price by a little bit less. Rather than following in a similar fashion, as it had done in the past, P&G responded with its own price cut. It matched the price decrease and kept the number of diapers per package the same. The result: P&G gained significant market share at the expense of Kimberly-Clark.

P&G eventually did reduce the number of diapers in the package to match the change by Kimberly-Clark, but not for several months after the initial move. The damage had been done, however. P&G reported significantly increased earnings for the first quarter of 2003, while

Kimberly-Clark had to admit to the sluggish growth of its own sales. P&G increased its share of the disposable diaper market by close to 3 percentage points while Kimberly-Clark's share fell by 1 percentage point during the year. As P&G's CEO A.G. Lafley states, "When times are tough, you build share."[22]

This example of the contest between Kimberly-Clark and P&G in the market for disposable diapers is reminiscent of the derivation of the kinked demand curve in our model of non-collusive oligopoly. When one firm raises price and another firm keeps price unchanged, the result is a significant reduction in sales and profits for the firm that raises price. The reason is that the demand facing the firm is very elastic for prices above the current market price because consumers can easily purchase a similar product from a competitor at the existing price. This result is essentially what occurred when Kimberly-Clark increased the per unit price of Huggies diapers while P&G reduced the per unit price of Pampers.

Summary of Key Points

1. The four-firm concentration ratio and the Herfindahl-Hirschman Index (HHI) are two measures of industry concentration, but the HHI is the most exact measure.
2. Monopolistically competitive markets have low degrees of industry concentration and each firm has a small amount of market power, which derives from product differentiation.
3. Monopolistically competitive firms may experience short run profits or losses, but in the

long run economic profits are zero due to the lack of entry barriers.

4. Oligopolistic markets have high degrees of industry concentration, and each firm behaves in a strategic fashion with respect to its rivals.
5. The kinked demand curve model explains price rigidity in non-collusive oligopoly markets.
6. Cartels arise in collusive oligopoly markets and function much like pure monopolies, but such agreements are difficult to maintain for a variety of reasons.
7. The price leadership model is used to explain how markets are divided between a large, dominant firm and many small competitors.
8. The outcome of a game is a dominant equilibrium when both players have dominant strategies (i.e., they always choose the same strategy as their best response).
9. The outcome of a game is a Nash equilibrium when each player offers her best response given the other player's choice of strategy.
10. In post-Keynesian pricing theory, oligopolistic firms adjust output in response to short run changes in demand and only change price when normal costs change or when the markup is adjusted to ensure sufficient funds for investment projects.

List of Key Terms

Monopolistic Competition

Oligopoly

Market share

Four-firm concentration ratio

Herfindahl-Hirschman Index (HHI)

Unconcentrated

Moderate degree of concentration

High degree of concentration

Excess capacity theorem of monopolistic competition

Game theory

Duopoly market

Supply price

Kinked demand curve model of non-collusive oligopoly

Cartel

Price leadership model

Players

Simultaneous games

Strategies

Payoff matrix

Outcomes

Payoffs

Dominant strategy

Best response

Dominant equilibrium

Suboptimal outcome

Nash equilibrium

Flexprice markets

Fixprice markets

Markup

Problems for Review

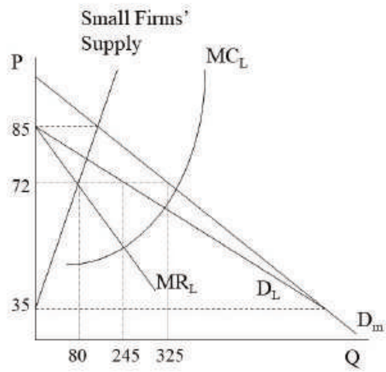
1. Suppose in an industry with annual sales of \$8 million, Firm 1 has annual sales of \$1.4 million. What is Firm 1's annual market share (S_1)?

2. Suppose an industry is composed of five firms with the following market shares: 17, 25, 8, 27, and 23.

- Calculate the industry's four-firm concentration ratio.
- Calculate the Herfindahl-Hirschman index for the industry.
- Would the United States Department of Justice designate the industry as unconcentrated, moderately concentrated, or highly concentrated?

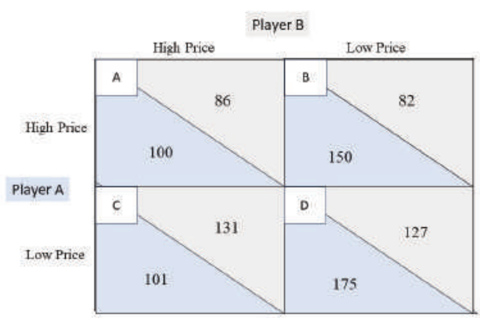
3. Suppose the price leader sets the profit-maximizing price as represented in Figure 10.15. How much output does the dominant firm produce and how much output do the small firms produce?

Figure 10.15: Problem #3



4. Consider the game in Figure 10.16:

Figure 10.16: Problem #4

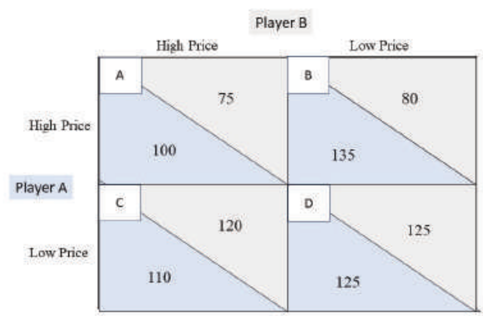


- Which players have dominant strategies in the game?
- What is the outcome of the game?

- Is the outcome of the game a dominant equilibrium?
- Is it a Nash equilibrium?

5. Consider the game in Figure 10.17:

Figure 10.17: Problem #5



- Which players have dominant strategies in the game?
- What is the outcome of the game?
- Is the outcome of the game a dominant equilibrium?
- Is it a Nash equilibrium?

Notes

1. Antitrust Division. “Horizontal Merger Guidelines (08/19/2010): 5.3 Market Concentration.” The United States Department of Justice. Web. Updated June 25, 2015. Accessed

on May 11, 2018. https://www.justice.gov/atr/horizontal-merger-guidelines-08192010#5c

2. All the figures for unconcentrated, moderately concentrated, and highly concentrated industries are from the United States Census Bureau. “Manufacturing: Subject Series – Concentration Ratios: Share of Value Added Accounted for by the 4, 8, 20, and 50 Largest Companies for Industries: 2002.” 2002 Economic Census of the United States. Web. Accessed on May 11, 2018.
3. For practical purposes, economists frequently calculate the HHI using the top 50 or top 20 firms in the industry even though technically all firms’ market shares should be included in the calculation.
4. Reed, Brian. “Money in a Bottle: The Celebrity Scent Business.” National Public Radio. November 6, 2009.
5. Samuelson and Nordhaus (2001), p. 212, present this case in a rather different way but arrive at the same result.
6. See Sweezy, Paul. “Demand Under Conditions of Oligopoly.” *Journal of Political Economy*. Vol. 47, 1939. Pp. 568-573, which is also cited in Keat et al.’s (2013), p. 367, summary of the kinked demand model of oligopoly.
7. Similar lists of factors that influence cartel formation are provided in many neoclassical textbooks. See Keat et al. (2013), p. 391-392, and McConnell and Brue (2008), p. 459-460.
8. See Varian (1999), p. 476, for a similar graphical representation of the price leadership model.
9. McConnell and Brue (2008), p. 464, emphasize this point.
10. Kenyon (1978), p. 34.
11. Ibid. p. 34.
12. Ibid. p. 34.
13. Ibid. p. 34-35.

14. Ibid. p. 35.
15. Ibid. p. 41.
16. Ibid. p. 38-39.
17. Ibid. p. 37-38.
18. Ibid. p. 39-40.
19. Ibid. p. 40.
20. The model in this section is based on the model presented in Snowdon, et al. (1994), p. 370-372.
21. Snowdon et al. 1994: 372
22. Sarah Ellison. “In Lean Times, Big Companies Make a Grab for Market Share.” *The Wall Street Journal*. September 5, 2003.

CHAPTER 11

THEORIES OF THE LABOR MARKET

Goals and Objectives:

In this chapter, we will do the following:

1. *Describe* the neoclassical theory of the market for labor
2. *Explore* the neoclassical theory of monopsonistic labor markets
3. *Consider* whether neoclassical monopsony theory represents a theory of exploitation
4. *Analyze* the case of bilateral monopoly in the labor market
5. *Investigate* the Marxian theory of the market for labor-power
6. *Explain* how changes in the character of production influence prices in Marxian theory

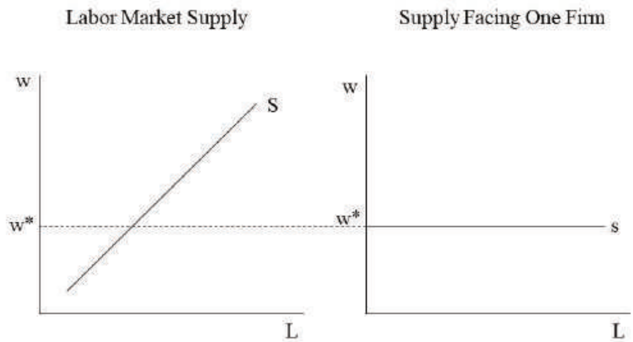
Up until this chapter, the focus has been almost exclusively on markets for goods and services sold to the final consumer. **Factor markets** (also referred to as **input markets** or **resource markets**) include the markets for labor, capital, and land. As the reader might expect, different schools of economic thought possess different theories of how these markets function. In this chapter, we will concentrate on the market for labor. We will take a close look at how the labor market operates from a microeconomic perspective, according to neoclassical economists and Marxian economists.

The Neoclassical Theory of the Demand for Labor

In neoclassical economic theory, product markets determine product prices and quantities exchanged. Similarly, neoclassical economists argue that labor markets determine wage rates and employment levels. The theory is essentially a story of supply and demand, much like the one we discussed regarding product markets. A sophisticated analysis underlies this story of supply and demand. This underlying story is developed at length in this section.

We begin with the assumption that the market supply of labor is upward sloping. That is, it is assumed that as the wage rate increases, the quantity supplied of labor rises as well, other factors held constant. Furthermore, it is assumed that the labor market is perfectly competitive such that each employer takes the market wage as given and so is a **wage-taker**. In other words, no single employer has any power to influence the wage that is paid. In this case, the labor supply curve facing the perfectly competitive firm is completely horizontal. This situation is depicted in Figure 11.1

Figure 11.1: The Labor Supply Curve Facing a Firm in a Perfectly Competitive Labor Market



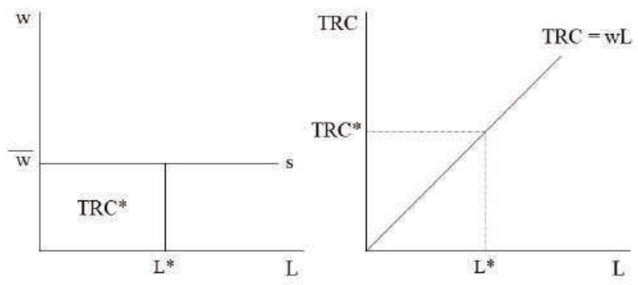
If an employer reduces the wage paid below the market wage even by a small amount, the quantity supplied of labor will fall to zero. That is, all workers will seek jobs from other employers who are offering the market wage. Similarly, the smallest increase in the wage above the market wage will lead to a sharp (infinite) increase in the quantity supplied of labor. In other words, the **wage elasticity of labor supply** facing a single employer is infinite in the case of a perfectly competitive labor market.

To understand how much labor an employer will hire to maximize its economic profit, we need to explore the implications of this wage-taking behavior for the firm's production costs. One concept that is important for this purpose is **total resource cost** (TRC). The TRC is the total cost of purchasing labor or the total wage bill. It is defined more precisely as follows:

$TRC = wL$

The relationship between the horizontal labor supply curve facing an employer and the employer's TRC curve is shown in Figure 11.2.

Figure 11.2: The Total Resource Cost Curve



As Figure 11.2 shows clearly, the TRC grows continuously as more labor is hired due to the constant wage rate that must be paid to each worker.

Furthermore, we can define **average resource cost** (ARC) as the average cost per worker hired. That is, if we were to spread out the total labor cost over the number of workers hired, then we would have the ARC. The ARC is defined precisely as follows, which it turns out can be reduced to the wage:

$ARC = \frac{TRC}{L} = \frac{wL}{L} = w$

This result should not be surprising. It simply means that, on average, the cost of a unit of labor is the wage. Because the wage is given, it follows that the wage must be the average cost of this resource.

Additionally, we can define the **marginal resource cost** (MRC) as the additional resource cost incurred with the purchase of an additional worker hired. Because the TRC grows by a constant amount equal to the wage rate with the purchase of each additional unit of labor, the MRC is the wage rate. It can be defined more exactly as follows:

$$MRC = \frac{\Delta TRC}{\Delta L} = w$$

The reader should notice that the MRC is equal to the slope of the TRC curve. The slope of the TRC curve, of course, is equal to the wage rate. Finally, it should be noted that because the ARC and the MRC are both equal to the wage rate, the ARC and MRC curves will be identical to the horizontal labor supply curve facing an employer. Figure 11.3 adds the ARC and MRC curves to a graph of the horizontal labor supply curve facing one employer.

Figure 11.3: The Average Resource Cost Curve and the Marginal Resource Cost Curve

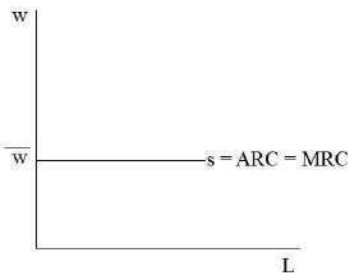


Table 11.1 provides a numerical example that includes calculations of TRC, ARC, and MRC.

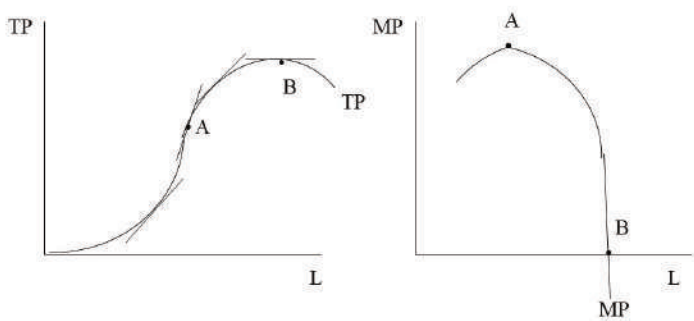
Table 11.1: Supply and Cost Information Facing a Perfectly Competitive Employer

L	w	TRC	MRC	ARC
0	10	0	--	--
1	10	10	10	10
2	10	20	10	10
3	10	30	10	10
4	10	40	10	10
5	10	50	10	10
6	10	60	10	10

As expected, the MRC and ARC are equal to the wage rate, and the TRC grows continuously with employment due to the constant wage rate.

The employer must consider the effect on revenue of hiring additional labor as well as the effect on cost. As a result, we need to introduce a new concept that neoclassical economists refer to as the **marginal revenue product** (MRP) of labor. To understand this concept, we need to return to the total product (TP) and marginal product (MP) curves that were first introduced in chapter 7. Figure 11.4 shows the graphs of the total product and marginal product curves.

Figure 11.4: The Total Product and Marginal Product Curves



The reader should recall that marginal product is simply the slope of the total product curve. In the short run, MP rises due to specialization and division of labor at low employment levels and then falls due to diminishing returns to labor at higher employment levels. The MRP refers to the additional revenue earned from the purchase of an additional unit of labor, which can be defined as follows:

$MRP = \frac{\Delta TR}{\Delta L}$

The MRP can be further expanded as follows:
$MRP = \frac{\Delta TR}{\Delta L} = \frac{\Delta TR}{\Delta Q} \cdot \frac{\Delta Q}{\Delta L} = MR \cdot MP$

In other words, the MRP is the mathematical product of the firm's marginal revenue and marginal product of labor. Finally, we learned in Chapter 8 that a perfectly competitive firm's marginal revenue is equal to the given market price. If we assume that this firm is a perfectly competitive producer, then the MRP can be written as follows:
$MRP = MR \cdot MP = P \cdot MP$

This result is very intuitive. The MRP is the additional revenue that a firm earns from hiring another unit of labor. When the additional unit of labor is purchased, it will produce some additional output. This additional output is the marginal product (MP). The additional output is then sold at the given market price (P) in a perfectly competitive market. The additional revenue generated is the MRP.

A closely related concept is the **average revenue product** (ARP) of labor. It is simply the total revenue per worker hired, which may be defined as follows:
$ARP = \frac{TR}{L}$

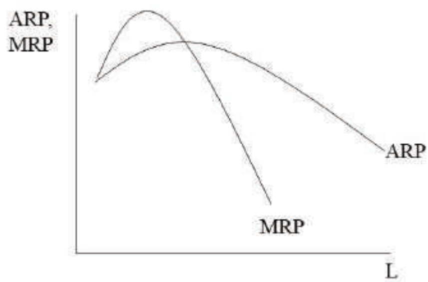
Figure 11.5 shows the calculation of the MRP and the ARP for a perfectly competitive firm that faces a constant market price of $2 per unit.

Figure 11.5: The Marginal Revenue Product (MRP) for a Perfectly Competitive Firm

L	TP	MP	P	TR	MRP	ARP
0	0	--	2	0	--	--
1	6	6	2	12	12	12
2	13	7	2	26	14	13
3	18	5	2	36	10	12
4	22	4	2	44	8	11
5	25	3	2	50	6	10
6	27	2	2	54	4	9
7	28	1	2	56	2	8

As Figure 11.5 shows, the MRP rises and then falls as employment rises. Because the MRP equals the product price times the marginal product of labor and the product price is constant, the MRP will rise due to the specialization of labor, but then it must fall because of the fall in marginal product. That is, the MRP falls due to diminishing returns to labor. A similar argument explains the shape of the ARP curve. Figure 11.6 provides a graph of the MRP and ARP curves.

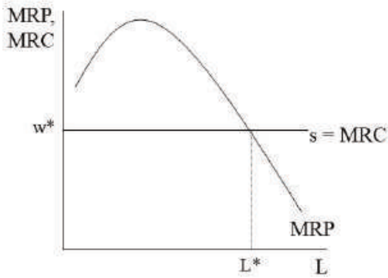
Figure 11.6: The Marginal Revenue Product Curve and the Average Revenue Product Curve for a Perfectly Competitive Firm



As with marginal product, both the ARP and the MRP rise initially due to labor specialization and both fall eventually due to diminishing returns to labor.

It is now possible to determine the profit-maximizing quantity of labor employed. Figure 11.7 places the MRC and the MRP on a single graph and in a single table.

Figure 11.7: A New Profit-Maximizing Rule:
MRP = MRC



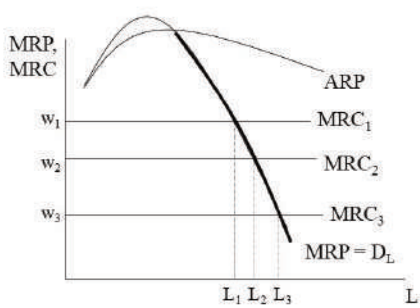
L	MRP	MRC
0	--	--
1	12	10
2	14	10
3	**10**	**10**
4	8	10
5	6	10
6	4	10

If a unit of labor generates more additional revenue for a firm than it adds to cost, the firm will increase its economic profits by purchasing that unit of labor. For example, the first unit of labor adds $12 to revenue and only $10 to cost. The purchase of that unit of labor will increase the firm's economic profits. As the firm continues to purchase more labor, the MRP eventually falls due to diminishing returns to labor and the MRC remains constant due to the given market wage. When 3 units of labor are purchased, the MRP and the MRC are equal at $10. In other words, the third unit of labor adds as much to revenue as it adds to cost. The firm is technically indifferent between purchasing and not purchasing this unit of labor. By convention, neoclassical economists argue that this unit will be purchased. Any units of labor beyond this point, however, will add more to cost than revenue and their purchase will reduce the firm's economic profits. The employer, therefore, will maximize economic profits where MRP equals MRC.

Furthermore, the perfectly competitive employer achieves maximum economic profits where MRP equals the wage rate.

This profit-maximizing rule for the labor market can be used to derive the **labor market demand curve**. It is only necessary to consider what happens to the quantity of labor employed at different wage rates, holding all other factors constant. As the wage rate decreases, the employer maximizes economic profits by choosing the quantity of labor such that w equals the MRP as shown in Figure 11.8.

Figure 11.8: The MRP Curve as the Labor Demand Curve



It is easy to see that as the wage rate falls, the quantity of labor demanded increases. In other words, the labor demand curve is downward sloping. Furthermore, because the quantity of labor demanded is determined at points of intersection between the wage rate and the

MRP, the MRP curve is the perfectly competitive employer's labor demand curve.

A second condition is necessary to ensure that economic profits are maximized in the short run. Specifically, it must be proven that the employer cannot earn a larger economic profit by shutting down in the short run. To prove this point, we return to the shut-down rule that we learned when we first discussed how a perfectly competitive firm maximizes economic profits in Chapter 8. In Chapter 8, it was shown that a perfectly competitive firm should only operate when price is at least as great as average variable cost (P ≥ AVC). We may derive a similar shut down rule for a perfectly competitive employer as follows:

$$P \geq AVC$$
$$P \cdot Q \geq AVC \cdot Q$$
$$TR \geq TVC$$
$$\frac{TR}{L} \geq \frac{TVC}{L}$$
$$\frac{TR}{L} \geq \frac{wL}{L}$$
$$ARP \geq w$$

In other words, the wage paid must be less than or equal to the ARP. Otherwise, the employer should shut down in the short run. Therefore, when we trace out the labor demand curve, the only relevant portion of the MRP curve is that part that is below the ARP curve, as shown in Figure 11.8. Because the MRP curve intersects the ARP curve at the maximum point on the ARP curve, the highest wage rate at which a positive amount of labor is demanded is the maximum ARP. If the wage rate rises above this point, then the employer will shut down and demand no labor.

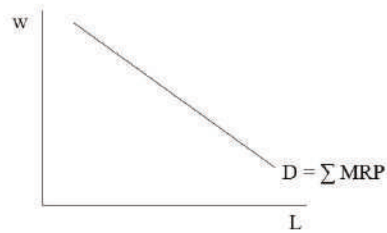
Neoclassical economists assert that the demand for labor (or any input) is a **derived demand**. That is, labor demand is derived from the demand for the product that the labor produces. For example, the demand for engineers depends on the demand for new construction. If firms invest in the construction of more bridges, dams, and skyscrapers, then they will need to hire more engineers. As we learned in Chapter 3, if the demand for a product rises, then the market price will increase, other factors held constant. The rise in the market price will cause the marginal revenue product of labor to increase because the output that workers produce can now be sold at a higher price. This change causes an outward shift of the MRP curve. Because the MRP curve is the employer's labor demand curve, the labor demand curve shifts outward. Therefore, a rise in the demand for a product causes a rise in the demand for the labor that produces it.

More generally, we can identify two changes that can shift the MRP curve and thus the labor demand curve. Recalling that the MRP is equal to the product price times the MP of labor, we can identify these two changes as follows:

1. Any factor that raises the price of the product will increase the MRP of labor and thus the demand for labor.
2. A change in production technology, or any other change that increases the marginal product of labor, will increase the MRP of labor and thus the demand for labor.

Now that we have derived the labor demand curve for a perfectly competitive employer, it is a short step to obtain the demand curve for the entire labor market. We simply use horizontal summation to aggregate the individual labor demand (or MRP) curves of many different employers. The downward sloping labor market demand curve that results from this aggregation process is shown in Figure 11.9.

Figure 11.9: The Market Demand for Labor



The Neoclassical Theory of the Supply of Labor

Neoclassical theorists have also developed a theory of labor supply. According to this theory, individual workers allocate their available time between working time and leisure time to maximize utility. In this theory, work is regarded as undesirable for its own sake, but it provides wage income that can be used to purchase commodities. Leisure time, on the other hand, is generally regarded as desirable. Because wage income is desired to acquire consumer goods, the wage rate represents the opportunity cost of one hour of leisure time. That is, by choosing to enjoy an hour of leisure time, a worker sacrifices the wage that could be earned.

We can use the modern theory of utility maximization to represent the problem facing the individual worker.[1] To begin, we consider the **time constraint** that the worker faces and represent this constraint in much the same way that we represented the budget constraint facing a consumer in Chapter 6. In this theory, the worker has a total amount of time (T) available each day for either work or leisure. T will generally be less than 24 hours because the worker is unavailable for either work or leisure during sleeping hours. The hours spent working (h) and the hours of leisure time (l) add up to the total time available in the day as shown below:

$$T = h + l$$

Furthermore, the daily income (Y) is equal to the wage rate (w) times the number of hours spent working (h) as shown below.

$$Y = wh$$

If we rearrange the above equation such that h = Y/w, then we can rewrite the total amount of time available in the day in the following way:

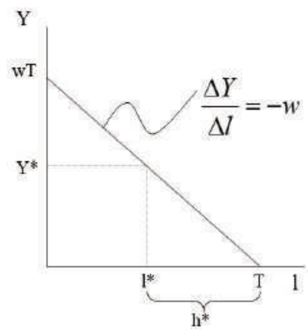
$$T = \frac{Y}{w} + l$$

Solving this equation for Y, we obtain the following result:

$$Y = wT - wl$$

It should be noted that w and T are unknown constants in this equation and Y and l are the only variables. If we graph this result as shown in Figure 11.10, then we obtain a clearer picture of the **income/leisure tradeoff** facing the individual worker.

Figure 11.10: The Time Constraint
(The Income/Leisure Tradeoff)

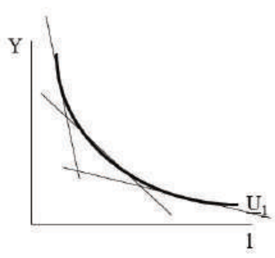


In Figure 11.10, it should be clear that if l is equal to zero, then Y = wT. This point represents the vertical intercept of the time constraint. In economic terms, if the worker chooses no leisure time and only chooses to work, then the maximum income that can be obtained is the wage rate times the total time available. It should also be clear that if Y is equal to zero, then l = T. This point represents the horizontal intercept of the time constraint. In economic terms, if the worker chooses to not work at all and thus earns no income, then the maximum leisure time is the total time available in the day.

Three other comments need to be made about the time constraint represented in Figure 11.10. If the worker chooses l* amount of leisure and earns Y* amount of income, then the hours of work (h*) can be represented as the difference between the horizontal intercept (T) and the amount of leisure chosen. Additionally, the slope of the line has a special significance. The slope (ΔY/Δl) is equal to the negative of the wage rate (–w). That is, a one hour increase in leisure time will lower the worker's income by an amount w. As stated previously, the opportunity cost of an hour of leisure time is the wage rate. Finally, a change in the amount of time available will shift the line, and a change in the wage rate will change the vertical intercept and the slope but leave the horizontal intercept unchanged.

Just as the individual consumer's preferences for goods may be represented using indifference curves, the individual worker's preferences for income and leisure may be represented using indifference curves as shown in Figure 11.11.

Figure 11.11: A Worker's Indifference Curve



In Chapter 6, we learned that the downward slope of an indifference curve indicates that the consumer is willing to trade off one good for another. Similarly, the downward slope of the indifference curve represented in Figure 11.11 indicates that the worker is willing to trade off income for leisure and vice versa. We also learned in Chapter 6 that the slope of the indifference curve is called the marginal rate of substitution (MRS) and that this slope becomes flatter as the individual moves along the indifference curve. The reason for the change in the slope is that as the worker obtains more leisure, her willingness to trade off additional income for an additional hour of leisure decreases. This diminishing marginal rate of substitution is somewhat like diminishing marginal utility. As explained in Chapter 6, however, diminishing MRS depends entirely on an ordinal notion of utility.

We can also rewrite the MRS as the negative ratio of the marginal utilities of leisure and income. Because the worker's utility remains the same all along the indifference curve, we can write the following equation:

$$\Delta TU = MU_Y \cdot \Delta Y + MU_l \cdot \Delta l = 0$$

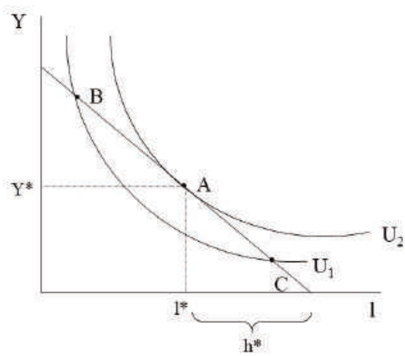
This equation states that the change in total utility as the worker moves along an indifference curve is equal to the product of the marginal utility of income (MU_Y) and the change in income (Y) *plus* the product of the marginal utility of leisure (MU_l) and the change in leisure (l). The entire sum is equal to zero because total utility remains constant along the indifference curve. Solving for the MRS generates the following result:

$$MRS = \frac{\Delta Y}{\Delta l} = -\frac{MU_l}{MU_Y}$$

As the worker moves to the right along the indifference curve, the amount of l that is chosen rises and the amount of Y that is chosen declines. As a result, the marginal utility of leisure declines relative to the marginal utility of income, implying diminishing MRS.

We can now represent the utility maximizing choice of the worker. In Figure 11.12, the worker maximizes utility at point A by choosing the amount of leisure (l*) and hours of work (h*) that yields an amount of income, Y*.

Figure 11.12: Worker Equilibrium: Utility Maximization

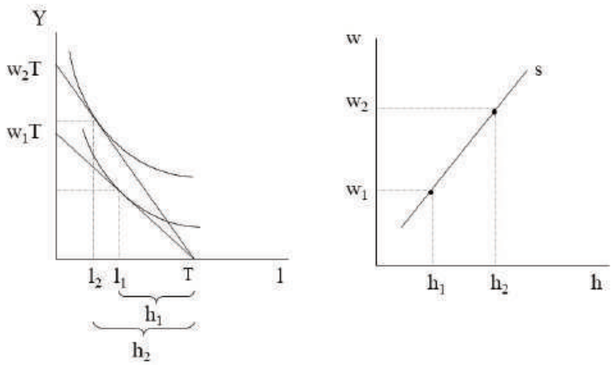


At point A, the indifference curve passing through point A is tangent to the time line representing the worker's time constraint. Because the slopes of these curves must be the same, the following condition must hold:

$$MRS = -\frac{MU_l}{MU_Y} = -w$$

It is now possible to derive the individual worker's labor supply curve using the utility maximizing framework that we have developed. Figure 11.13 shows what happens when the wage rate increases.

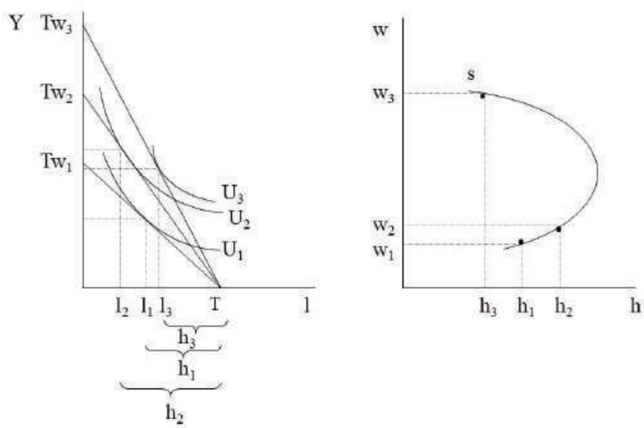
Figure 11.13: The Derivation of the Individual Worker's Labor Supply Curve



In Figure 11.13, the vertical intercept increases because the maximum possible income is now higher. Similarly, the slope increases (in absolute value) because the opportunity cost of leisure has increased with the higher wage rate. Because leisure has become more expensive to consume, the worker chooses to reduce the amount chosen from l_1 to l_2. The amount of work chosen correspondingly increases from h_1 to h_2. The quantity supplied of labor thus rises with the wage rate, which implies the upward sloping labor supply curve shown in the graph on the right in Figure 11.13.

On the other hand, it is possible that the worker will stop responding in this manner to the rise in the wage once the wage reaches a very high level. Suppose that the increase in the wage leads the worker to feel richer overall. As a result, the worker purchases more consumer goods but also decides to "purchase" more leisure time by working less. This situation is represented in Figure 11.14.

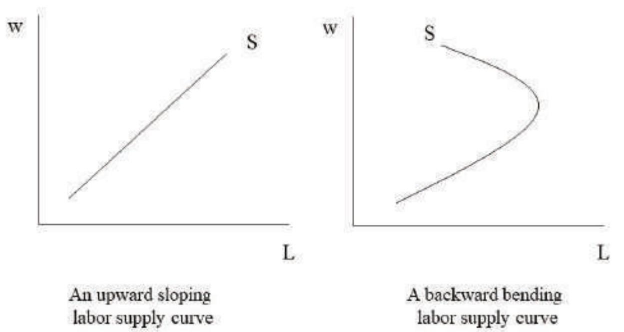
Figure 11.14: The Backward – Bending Labor Supply Curve



In Figure 11.14, the wage rises from w_1 to w_2, and the worker cuts back on leisure time (from l_1 to l_2) as leisure becomes costlier. Similarly, the hours worked increase from h_1 to h_2 as before. Once the wage rises to w_3, however, the worker increases leisure time from l_2 to l_3. A corresponding reduction in hours worked from h_2 to h_3 occurs. This drop in the number of hours worked as the wage increases is represented in the graph on the right in Figure 11.14 as a backward bending labor supply curve.

Our final step is to aggregate the individual labor supply curves of every worker in the labor market. As before, we can use horizontal summation to obtain the labor market supply curve. Figure 11.15 shows two possible examples of the labor market supply curve.

Figure 11.15: The Labor Market Supply Curve

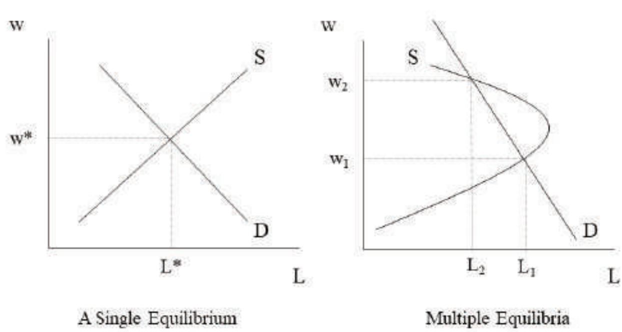


In the graph on the left in Figure 11.15, the labor market supply curve has the usual upward slope that we expect of a supply curve. As the wage rises, workers reduce their leisure time (which is more expensive) and work more to take advantage of the higher wage. The tendency to consume less leisure as the wage rises (other factors held constant) is referred to as the **substitution effect** in this context in the sense that the worker substitutes away from something that has become relatively more expensive to consume. In the graph on the right, however, workers respond to higher wages by eventually working less and consuming more of all goods, including leisure. The tendency to purchase more of all goods as one's income rises (other factors held constant) is referred to as the **income effect**. That is, the worker experiences a rise in real income and so decides to purchase more of everything. Although both effects are typically at work, whether an upward sloping or backward bending supply curve emerges depends on which effect is the stronger of the two. If the substitution effect dominates, then the labor supply curve will be upward sloping. If the income effect dominates, then the labor supply curve will be backward bending.

The Neoclassical Theory of Labor Market Equilibrium

Now that we have developed both the supply and demand sides of the labor market, we can bring them together to show how neoclassical economists explain the movement to equilibrium in these markets. Figure 11.16 shows two possible labor markets.

Figure 11.16: Labor Market Equilibrium



In the graph on the left, a single equilibrium outcome occurs. The labor market is cleared of shortages as wages increase, and it is cleared of surpluses as wages decrease.

Eventually, the market reaches an equilibrium wage rate and employment level. The market will remain at this point unless it is disturbed by a change in an external variable.

In the graph on the right, two different equilibrium outcomes are possible due to the backward bending supply curve, which causes a second intersection with the labor market demand curve. The lower equilibrium at w_1 and L_1 is the same as the one represented in the graph on the left. It is a **stable equilibrium** in the sense that a slightly higher or lower wage will lead to a surplus or shortage that will push the wage back in the direction of the equilibrium outcome. The upper equilibrium at w_2 and L_2, on the other hand, is rather different. If the wage falls below w_2 by a small amount, then it will continue to fall due to the surplus that exists. Similarly, if the wage rises above w_2 by a small amount, then it will continue to rise due to the shortage that exists. Because the wage tends to move further away from the equilibrium when pushed in either direction by a small amount, the equilibrium is an **unstable equilibrium**. The presence of an unstable equilibrium creates a risk of considerable market instability.

We have yet to mention the ideological significance of the neoclassical theory of the labor market. The neoclassical model of a perfectly competitive labor market reaches the conclusion that each worker is paid according to that worker's contribution to production. Earlier in this chapter, it was shown that a perfectly competitive employer achieves maximum economic profits when the MRP is equal to the wage rate (the MRC). This conclusion means that when the labor market reaches equilibrium each worker will receive a wage that is equal to the worker's contribution to the firm's revenue. From a purely ideological perspective, this result is a very powerful one. It means that workers are not exploited as Marxian economists assert. They draw from the social product an amount that is exactly equal to what they contribute. The **marginal productivity theory of income distribution** is implicitly a theory of distributive justice. That is, people receive what they deserve to receive. What they deserve to receive stems from their productive contributions.

The theory has been criticized for a variety of reasons. One objection is that inequality may have its own undesirable social and economic consequences and that payment according to marginal revenue product might lead to extreme levels of inequality. A second objection is that the relationship between social classes (e.g., workers and capitalists) plays no role in the analysis as it does in Marxian economics. Due to the assumption of perfect competition, no employer or worker has any market power. The fact that some own the means of production whiles others lack means of production is given no significance in the model. Finally, the assumption of perfect competition in the labor market is one that opponents of the theory have sharply criticized. As we will see in the next section, when the assumption of perfect competition is dropped, the door to a neoclassical theory of exploitation is suddenly thrown open.

A Neoclassical Theory of Exploitation?

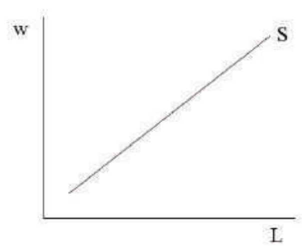
If we drop the assumption of perfect competition in the labor market, then how will the neoclassical analysis of the labor market change? In this section, we consider the case of imperfectly competitive labor markets. We will consider the case of a single employer, also referred to as a **monopsony** employer. A monopsony exists in a market when only a single buyer exists. In the labor market, the employers are on the buyers' side of the market. Therefore, a monopsonistic labor market is a market with only a single buyer of labor.

Pure monopsonies, just like pure monopolies, are not very common, but sometimes firms approach monopsony status in certain markets. For example, Wal-Mart has been accused of acting as a monopsonist in certain markets in which it buys goods from suppliers. In those cases, Wal-Mart is by far the largest, or may be the only, buyer of a product from its suppliers. In the defense industry, the U.S. government may be the only purchaser of advanced weaponry from firms that produce such products. Monopsony employers, on the other hand, have existed in company towns like General Motors in Flint, Michigan or Carnegie Steel in Homestead, Pennsylvania. In company towns, the people may have limited mobility and so they either work for the dominant employer, or they do not work at all.

As with any neoclassical model, we will start by identifying the model's main assumptions. We will assume that a single firm exists that is the sole buyer of labor. Furthermore, it is assumed that workers cannot easily move to a new location. Because of these conditions, the monopsonist has the power to set the market wage. That is, the monopsonist has market power, much like the monopolist possessed market power (i.e., the power to set the market price of its product).

Unlike the perfectly competitive employer who faces a horizontal labor supply curve, the monopsonist faces an upward sloping labor supply curve, as shown in Figure 11.17.

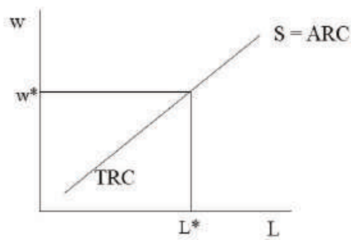
Figure 11.17: The Monopsonist and Supply



The reason for the upward slope of the labor supply curve facing the monopsonist is that the monopsonist faces the entire labor market supply curve, which is upward sloping. In general, as wages rise, more workers enter the labor market.

Just like the perfectly competitive employer, the monopsonist possesses an average resource cost (ARC) curve, as shown in Figure 11.18.

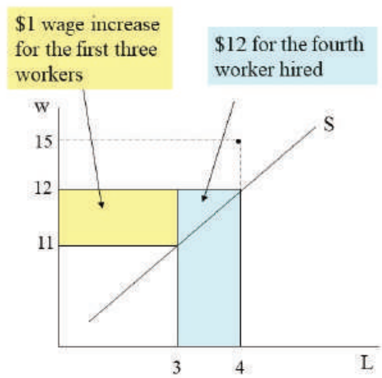
Figure 11.18: The Average Resource Cost Curve of the Monopsonist



It is assumed that the monopsonist establishes the same wage for each worker hired. Therefore, the ARC for the monopsonist is equal to the wage. The derivation of this result is the same as that which was used in the perfectly competitive case earlier in this chapter. Also, because the ARC is equal to the wage at each employment level, the ARC curve and the supply curve facing the monopsonist are one and the same.

The more interesting distinction between this labor market structure and the perfectly competitive one relates to the nature of the marginal resource cost (MRC) curve. Figure 11.19 shows an example of how to calculate the MRC for a monopsony employer.

Figure 11.19: Marginal Resource Cost for a Monopsonist



As the wage rises from \$11 per unit to \$12 per unit, the quantity of labor supplied increases from 3 to 4 units. The MRC may be calculated by dividing the change in TRC by the change in L as follows:

$$MRC = \frac{\Delta TRC}{\Delta L} = \frac{(12)(4)-(11)(3)}{4-3} = \frac{48-33}{1} = \$15\ per\ unit\ of\ labor$$

It is possible to obtain this result in another way that is more intuitive. When the wage rate is increased from \$11 to \$12 per unit, an additional worker enters the market. How much does this increase in the wage add to cost? The additional worker is paid \$12, but the reader should notice that the three workers, who were receiving \$11 each, now receive \$1 raises. Therefore, the total resource cost rises by \$12 plus \$3 or \$15. This manner of proceeding is helpful in terms of understanding why the addition to total resource cost exceeds the wage paid. As the reader can observe, in Figure 11.19 the MRC of \$15 is above the wage of \$12. In general, the MRC will exceed the wage because when the wage rises to encourage another worker to enter the market, the TRC rises both because of the wage paid to the new worker hired but also because each of the existing workers must receive a wage increase. The reader might notice the similarity between this analysis and the analysis of pure monopoly. In the case of pure monopoly, MR falls faster than price because when the price is cut to sell another unit, the price must also be cut on all the other units previously sold at the higher price.

Because the MRC exceeds the wage, the MRC curve will rise more quickly than the labor market supply curve facing the firm. Therefore, we obtain the result shown in Figure 11.20.

Figure 11.20: The Marginal Resource Cost Curve for a Monopsonist

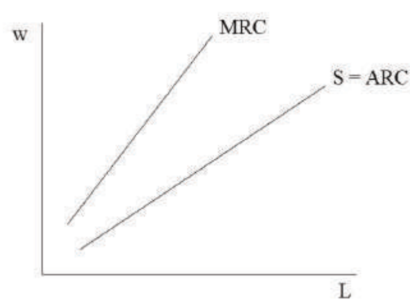


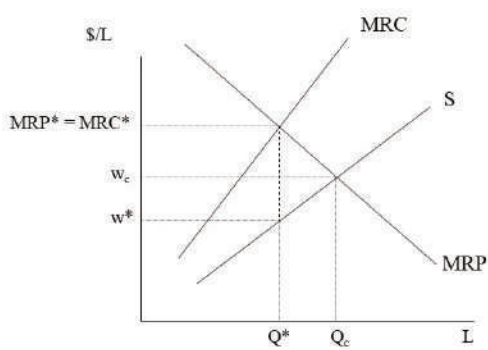
Table 11.2 provides an example to demonstrate how to calculate TRC, ARC, and MRC when only given information about the labor market supply curve facing the monopsony employer.

Table 11.2: Supply and Cost Information Facing a Monopsony Employer

L	w	TRC	MRC	ARC
0	8	0	--	--
1	9	9	9	9
2	10	20	11	10
3	11	33	13	11
4	12	48	15	12
5	13	65	17	13
6	14	84	19	14

We have now fully developed the cost structure of the monopsonist and can proceed to the profit-maximizing choice of the firm. Figure 11.21 shows how the monopsony employer will set the wage to maximize its economic profit.

Figure 11.21: Equilibrium in a Monopsonistic Labor Market



To maximize its economic profit, the monopsonist will hire labor up to the point where MRP equals MRC. This profit-maximizing rule in the factor market applies to the monopsony employer just as it applies to the perfectly competitive employer. Following this rule, the monopsonist will hire Q* units of labor. To encourage exactly this amount of labor to be supplied, however, the monopsonist must set the wage at w*. Only at w* is the wage at the level that is necessary to call forth Q* units of labor into the labor market. The crucial point to notice is that w* is significantly below the MRP, which means that the last (marginal) worker hired is paid a wage that is below the revenue that the worker generates for the employer. That is, the monopsonist forces wages down to make a greater economic profit.

This result (w* < MRP*) suggests that the worker is being exploited in an economic sense. For this outcome to be obtained within a neoclassical economic model is rather unusual. Such findings open the door to criticisms of unregulated market activity, which neoclassical economists generally favor. Neoclassical economists frequently respond to this finding by arguing that this case is an extreme one in which a single employer dominates the labor market. Typically, competition from other employers will drive wages up. Furthermore, even if this situation exists, neoclassical economists argue that the *degree of exploitation* as represented by the difference between the MRP and the wage (the dashed vertical line in Figure 11.21) is likely to be small or at least small enough that government efforts to correct this situation will lead to even worse economic consequences.[2]

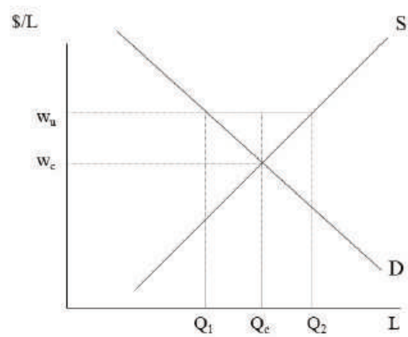
It is also possible to compare the monopsonistic equilibrium outcome and the perfectly competitive outcome using the graph in Figure 11.21. Earlier in this chapter, it was explained that the perfectly competitive equilibrium outcome in the labor market occurs where supply and demand intersect. If we assume that the MRP curve of the monopsony employer would be the same as the sum of the MRP curves of many perfectly competitive employers (if this market was perfectly competitive), then we can find the perfectly competitive equilibrium at the intersection of the labor market supply curve and the MRP curve. That is, the MRP curve represents the labor market demand curve and so its intersection with the labor market supply curve represents the competitive equilibrium. In the perfectly competitive equilibrium, w_C represents the equilibrium wage rate and Q_C represents the quantity of labor that the firm will hire at that wage. Because w* is less than w_C, it is easy to see that the monopsony firm reduces the wage to a level that is below what would be paid in a perfectly competitive labor market. Furthermore, because Q* is less than Q_C, it is also easy to see that the monopsony firm reduces overall employment below what would exist in a perfectly competitive labor market. The reduction of employment below the perfectly competitive level represents a loss of efficiency brought on by the monopsonist's pursuit of maximum economic profits.

The Economic Consequences of Labor Union Activity

In Chapter 3, the concept of a price floor was introduced. A price floor establishes a legal minimum price in a market. The price is permitted to rise above a price floor, but the price cannot fall below the price floor. **Industrial unions** are organizations that attempt to organize all the workers in an industry and then negotiate industry-wide wage floors for their members. Working hours and working conditions are other key points for negotiation. **Craft unions** have similar aims, but they only organize the workers who share a common skill or trade, such as carpentry, ironworking, or masonry. Unions possess market power on the sellers' side of the labor market. That is, a union is (sometimes) the sole seller of labor in a market, just as the monopsonist is the sole buyer on the buyers' side of the market.

A price floor that a labor union might negotiate is just a minimum wage. If the labor market is perfectly competitive, then this situation can be represented as in Figure 11.22.

Figure 11.22: An Industrial Union in a Market with Many Employers



In Figure 11.22, the perfectly competitive equilibrium occurs at w_c and Q_c, but if the labor union negotiates a wage floor of w_u, then the wage will not be permitted to fall below this level. The quantity supplied of Q_2 will exceed the quantity demanded of Q_1, and a surplus will exist. That is, unemployment will exist, and it will persist while the wage floor is in effect and no external changes occur. It is worth noting, however, that the unemployed may be divided into two different categories in this case. First, some workers lose their jobs due to the wage being pushed up above the competitive equilibrium wage. The number of workers that loses jobs is equal to $Q_c - Q_1$. Additionally, other workers enter the labor market precisely because the wage has been pushed up above the competitive equilibrium wage. The number of workers who enter the labor market only to become unemployed is $Q_2 - Q_c$. Overall, neoclassical economists condemn union activity because it reduces the overall amount of employment and causes inefficiency.

On the other hand, if each side of the market is dominated by a single participant, then the results are quite different. Suppose, for example, that an industrial union faces a monopsonistic employer in the labor market. That is, a single seller of labor confronts a single buyer of labor. This market structure is referred to as **bilateral monopoly** and is depicted in Figure 11.23.

Figure 11.23: The Bilateral Monopoly Model

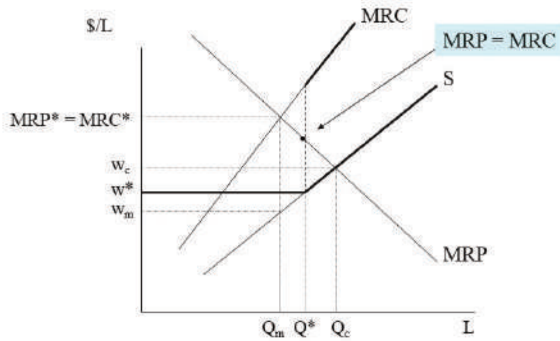


Figure 11.23 shows the monopsony outcome where the wage is w_m and employment is Q_m. It also shows the perfectly competitive labor market outcome where the wage is w_c and employment is Q_c. Let's assume, however, that a union negotiates a wage of w* with this monopsony employer. In this case, the labor market supply curve becomes perfectly horizontal for every employment level up to the original supply curve. The reason is that the workers who would have entered the labor market at lower wage rates previously are now paid the union wage. Once we reach the original supply curve, however, the wage must rise to encourage more workers to enter the market. The supply curve thus has a kink in it at Q*.

To obtain the MRC curve, it is necessary to use the information given on the supply curve. When the labor supply curve facing the firm is horizontal, as it is in the case of a perfectly competitive market, then the MRC curve is horizontal as well and identical to the supply curve. Therefore, the MRC will be the same as the labor market supply curve up to the kink. For employment levels beyond the kink, however, the MRC corresponding to the upward sloping supply curve applies. As a result, the MRC curve is horizontal up until the kink in the supply curve, then a vertical gap exists until we reach the upward sloping MRC curve, after which point the MRC curve becomes upward sloping as before.

To find the profit-maximizing outcome in the bilateral monopoly model, we only need to equate MRP and MRC. In Figure 11.23, the MRP curve intersects MRC somewhere in the gap in the MRC curve. This intersection gives us the profit-maximizing employment level of Q*. It also gives us the profit-maximizing wage. To call forth Q* amount of labor, the wage rate that must be set is w*, which is directly above Q* at the kink in the supply curve.

What we observe in this case is that the wage rate is higher than what the monopsonist would set in the absence of a union. The employment level is higher as well. Furthermore, the gap between the MRP and w is smaller and so the degree of exploitation is lower. On the other hand, it is also the case that the wage rate is lower than the perfectly competitive wage, and the employment level is lower than the perfectly competitive employment level. Still, if the labor union could negotiate an even higher minimum wage, then it would approach or even match the perfectly competitive outcome. If the union negotiates a wage that is higher than w_c, however, then unemployment will result as in the perfectly competitive model. The reader might try to verify this result graphically. In general, however, whether the wage that is negotiated is closer to the pure monopsony wage or closer to the perfectly competitive wage will depend on the relative bargaining strength of the monopsonist and the labor union. A relatively strong labor union will negotiate wages that are closer to the perfectly competitive result. A relatively weaker labor union will negotiate wages that are closer to the pure monopsony result.

A situation of bilateral monopoly like the one we have been discussing occurred in 1892 in Homestead, Pennsylvania. In a very famous strike, the Amalgamated Association of Iron and Steel Workers struck against the Carnegie Steel Company. The craft union had a large degree of monopoly power at the time, and Carnegie Steel was the only major employer in the entire town. The union's goals were to negotiate a minimum wage and to establish a June expiration date (rather than a January expiration date) for the new three-year contract. The union wanted a summer rather than a winter expiration date because if a strike became necessary during contract negotiations, the workers could hold out much better in the summer than in the winter. In this case, the company and the union were not able to arrive at an easy solution. A bitter strike ensued involving a battle between striking steelworkers and Pinkerton guards. Eventually, the Pennsylvania Governor ordered the state guard to force an end to the strike. Abstract models can teach us a great deal, but they often cannot capture the intensity of real life struggles.[3]

The Marxian Theory of the Market for Labor-Power

Now that we have studied the neoclassical theory of the labor market in considerable detail, we can more easily contrast it with the Marxian theory of the market for labor-power. The reader should recall that the commodity that workers sell to capitalists is labor-power as opposed to labor. In Marxian theory, labor refers to the act of working itself, whereas labor-power refers to the ability of a worker to perform labor for a given amount of time, which is sold as a commodity.

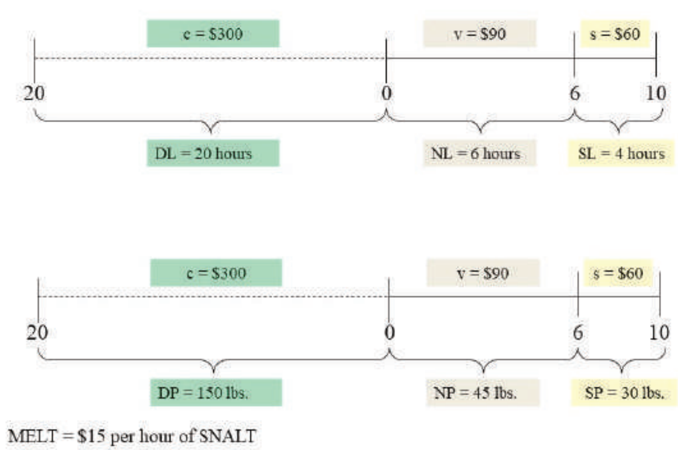
In Chapter 4, it was shown how the value of labor-power is determined in Marxian theory. Marx provided a formula for calculating the value of a day's labor-power. As the reader will recall, that calculation requires adding up all the values of all the means of subsistence that a worker requires in the year to produce and reproduce her labor-power (according to a culturally determined norm) and then dividing that value by the number of days in the year. If the social estimation of what a worker requires for the production and reproduction of labor-power changes, then the value of labor-power will change as well. Additionally, if the values of the required means of subsistence change, then the value of labor-power may change as well. The *price* of labor-power, which is what is paid for labor-power, may diverge from the value of labor-power at times. In the second part of this book, it is explained why the price of labor-power never diverges very much from the value of labor-power.

Our primary interest in this chapter, however, is to understand how changes in the capitalist production process can be analyzed from a Marxian perspective. This discussion draws heavily upon Marx's treatment of the subject in chapter 17 of volume 1 of *Capital*. In the remainder of this chapter, we will consider how productivity changes are treated in Marxian theory. We will also consider two aspects of capitalist production that are not given much attention in neoclassical theory, namely changes in the length of the working day and changes in the intensity of the labor process. As will be shown, the value of labor-power has an important role to play in the analysis.

Changes in the Productivity of Labor

In neoclassical microeconomic theory, an increase in the price of a firm's product raises the marginal revenue product of labor and thus labor demand. The causal claim is that price increases lead to increases in marginal revenue productivity. In Marxian economic theory, on the other hand, the causal chain runs in the reverse direction. That is, a rise in labor productivity typically leads to a reduction in prices. To see why, we need to return to our working day diagrams from Chapter 4. Figure 11.24 shows the three ways that we may express the value produced in one day in a specific industry.

Figure 11.24: Three Expressions of the Value Produced in a Day (in a specific industry)



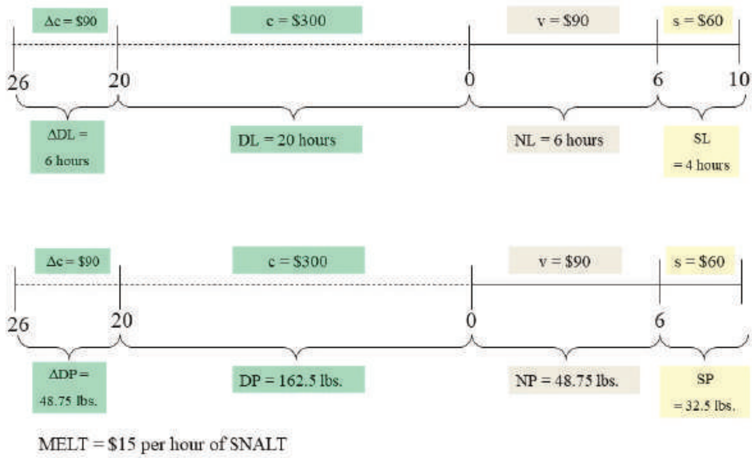
In this example, the capitalist advances constant capital (c) of \$300 for means of production and variable capital (v) of \$90 for labor-power. The worker works a 10-hour day. Given a monetary expression of labor time (MELT) of \$15 per hour, the \$90 of variable capital may be converted into 6 hours of necessary labor (NL). The remainder of the workday then consists of 4 hours of surplus labor (SL), which may be converted into \$60 of surplus value (s) using the MELT. The constant capital of \$300 may also be converted into its dead labor (DL) equivalent of 20 hours using the MELT. If we assume that the worker produces a total product (TP) 225 lbs. of sugar during the workday, then we can also calculate the individual value of a pound of sugar. All we need to do is divide the total value of the day's product by the total product. That is, the price (= value) can be calculated as follows:

$$p = \frac{c+v+s}{TP} = \frac{\$300+\$90+\$60}{225\ lbs.} = \frac{\$450}{225\ lbs.} = \$2\ per\ lb.$$

By dividing c, v, and s by the price of a pound of sugar, we can calculate the dead product (DP) of 150 lbs., the necessary product (NP) of 45 lbs., and the surplus product (SP) of 30 lbs., respectively.

Now that we have reviewed these basic aspects of Marxian economics, we can consider the effects of a change in labor productivity. Unlike in the neoclassical theory we considered earlier in this chapter, it matters a great deal whether the productivity change occurs in an industry that produces means of subsistence for workers (so-called wage goods industries) or in other industries that produce goods that workers do not typically consume. Let's first assume that a productivity increase occurs in an industry that is not a wage goods industry. This situation is depicted in Figure 11.25.

Figure 11.25: A 30% Productivity Increase
(in this industry but not in the wage goods industry)



In Figure 11.25, a 30% productivity increase is assumed. What this change means is that a worker can transform 30% more means of production (as reflected in a 30% rise in constant capital) into 30% more finished product in the same 10-hour workday as previously. That is, it is assumed that the worker produces more in the same 10 hours while working at the same level of intensity as previously. Indeed, this change represents a pure productivity increase. In this example, the additional $90 of constant capital (Δc) is used to purchase means of production representing 6 hours of additional dead labor (ΔDL). Similarly, the additional sugar produced may be considered an addition to the total product (ΔTP) of 67.5 lbs., which is a 30% increase. The new value of labor-power and the newly created value are not affected at all in this case. The price of sugar, however, is affected as can be observed in the following calculation:

$$p = \frac{c+\Delta c+v+s}{TP+\Delta TP} = \frac{\$300+\$90+\$90+\$60}{225\ lbs.+67.5\ lbs.} = \frac{\$540}{292.5\ lbs.} \approx \$1.85\ per\ unit$$

By dividing each monetary magnitude in Figure 11.25, we can calculate the new values for the surplus product, the necessary product, the dead product, and the change in dead product, as shown in Figure 11.25. To carry out these calculations, the exact figure for the price was used. As a result, when we add together each product figure, we obtain the new total product for the day of 292.5 lbs., which represents a 30% increase in production. The price of sugar, therefore, falls when labor productivity rises. By contrast, we would expect a productivity decline to increase the price of sugar.

The other possibility we should consider is a productivity change that occurs in a wage goods industry but *not* in the industry that we are considering. If productivity rises in a wage goods industry, then this change will have a direct impact on the value of labor-power. By reducing the value of the means of subsistence that the worker requires, the commodity labor-power becomes less valuable. If the price of labor-power falls in line with the drop in the value of labor-power, then this change will lead to a re-division of the workday in the industry that we are considering. This situation is depicted in Figure 11.26.

Figure 11.26: A Productivity Increase
(in the wage goods industry but not in this industry)

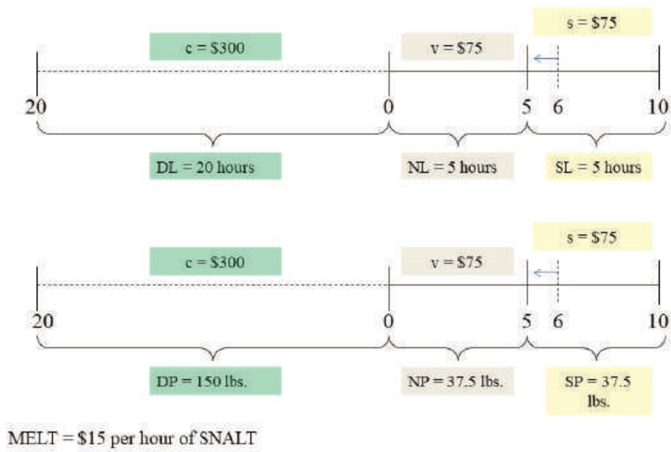


Figure 11.26 represents a situation in which the labor embodied in the required means of subsistence for the day falls to 5 hours of SNALT. With the necessary labor at 5 hours, the variable capital declines to $75 (given the MELT of $15/hour). In a similar fashion, the surplus labor rises from 4 hours to 5 hours (given the 10-hour workday), and the surplus value produced rises from $60 to $75. The constant capital advanced remains unaffected by this change in labor productivity in the wage goods sector. Because the total value of the day's product remains at the same level of $450 and the total amount of sugar produced remains unchanged at 225 lbs. of sugar, the price of sugar is not affected at all.

In Figure 11.26, the fall in the value of labor-power simply leads to a change in the distribution of the new value created. Aside from that change, production levels in this industry remain the same. Notice that workers receive a smaller money wage, but they can purchase the same quantity of means of subsistence as previously. Their absolute standard of living remains the same. Capitalists are the sole beneficiaries of the productivity increase in this case. It is possible that a struggle may develop between capitalists and workers over the division of the new value created. If labor unions are relatively strong, then the price of labor-power might rise above its new value (but perhaps not as high as the previous value of labor-power). In that case, the workers enjoy a higher standard of living, as they can purchase more means of subsistence than previously. At the same time, the capitalists extract more surplus value from the workers, and workers become poorer relative to capitalists. This possibility is interesting because it reveals that Marx's theory is consistent with rising real standards of living for workers even as inequality worsens.

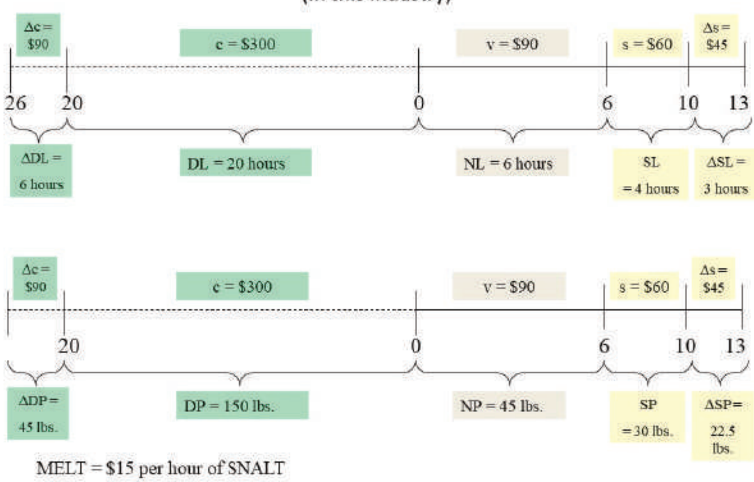
Of course, the one situation we have not considered is a productivity increase in a wage goods industry and the consequences of that change for the wage goods industry itself. This case would combine the two examples we have considered. That is, prices would fall in the wage goods industry and a part of the new value created would be redistributed from workers to capitalists as the value of labor-power declines. Although it is possible, it is not necessary to create a diagram for this case since it would simply reproduce the results we have already obtained in the previous two cases.

Changes in the Length of the Working Day

The next change we need to consider is a change in the length of the working day. Unlike in neoclassical theory where the worker decides how to allocate her time between work and leisure to maximize utility, in Marxian theory, capitalists tell workers what the length of the working day is. In the absence of a union, they either accept those terms or they seek work elsewhere.

Figure 11.27 represents a situation in which the working day has been extended.

Figure 11.27: A 30% Extension of the Working Day (in this industry)



In Figure 11.27, the working day is extended by 30% or 3 hours. Because workers must have means of production with which to work, this extension necessitates a 30% increase of $90 in the amount of constant capital advanced. The total product produced in the day subsequently rises by 30% or 67.5 lbs. The consequence of this increase in the length of the workday is an increase in the surplus value produced, but it has no effect on the price of sugar. The new (unchanged) price of sugar may be calculated as follows:

$$p = \frac{c+\Delta c+v+s+\Delta s}{TP+\Delta TP} = \frac{\$300+\$90+\$90+\$60+\$45}{225\ lbs.+67.5\ lbs.} = \frac{\$585}{292.5\ lbs.} = \$2\ per\ unit$$

The new dead product of 45 lbs. may be calculated simply by dividing the new constant capital advanced of $90 by this price.

The main consequence of an extension of the working day is an increase in the degree of exploitation. The value of labor-power is typically unaffected by such a change. Marx, however, did argue that the additional wear and tear that labor-power experiences due to this extension may increase the value of labor-power. That is, the means of subsistence necessary to make the production and reproduction of labor-power each day may rise due to, for example, an increased need for medical care. Beyond a certain point, however, no increase in the means of subsistence can compensate for the deterioration of the worker's health due to endless drudgery.

Additionally, if the value of labor-power remains unchanged even with a lengthening of the workday, it is possible that its price may increase above its value. That is, a struggle between workers and capitalists over the new value created might occur. Depending on the relative strength of the one versus the other, workers or capitalists may end up appropriating a larger portion of the newly created value as wages or surplus value, respectively.

Changes in the Intensity of Labor

The final change that we will consider is a change in the intensity of labor that occurs in a single industry but not across all industries simultaneously. For example, suppose that the intensity of the labor process increases above the social norm that exists in other industries. In this case, even with the same number of hours in the workday, the worker will create an even larger amount of new value than previously. The reason is that one hour of SNALT is not necessarily the same as one hour of clock time. If the intensity of labor rises above what is considered the social norm in a specific society, then one hour of clock time might be consistent with more than one hour of SNALT. This situation is depicted in Figure 11.28.

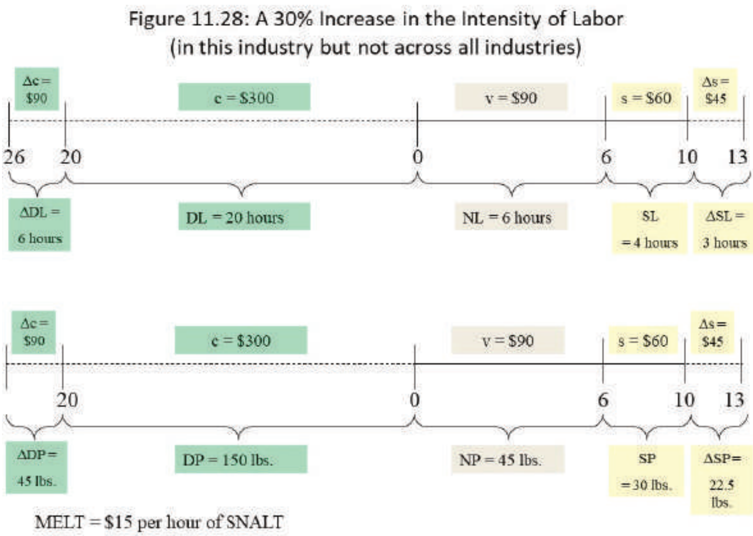


Figure 11.28 is almost identical to Figure 11.27, which depicted an increase in the length of the working day. The only difference is that the 3 hours of additional surplus labor do not occur because of an increase in the length of the workday. Instead, it is the result of 30% more work being performed within the span of the 10-hour workday. For this reason, the portion of the timeline that shows an extension of 3 hours is a dashed line rather than a whole line, as was the case in Figure 11.27. That is, the

increase in labor intensity leads to the incorporation of more SNALT in the final product and a greater value of the final product, but these additions are like the 30% increase in dead labor and constant capital advanced in that they are not part of the working day proper. On the other hand, these changes do represent new value created, and in that sense, they are very different from the contribution that the additional constant capital makes to the final product. In this case, because the value of the final product and the physical product both rise by 30%, the price of sugar remains unchanged. This result is to be expected because the numerical changes are identical to those obtained from an extension of the workday.

As in the case of an extension of the working day, the value of labor-power may rise due to its more rapid deterioration. Workers are not working longer hours, but they are working harder, which may impact their health. The same limits to compensating workers with a higher wage that apply in the case of the extension of the working day should also be expected to apply in this case. As before, even if the value of labor-power does not rise, workers might push for an increase in the price of labor-power as they struggle to win a portion of the newly produced value that their more intense labor has made possible.

The amount of new value created due to the intensification of the labor process is directly related to the extent of the divergence between the intensity of labor in this industry and the social norm. A general change in the intensity of labor across all industries, however, that alters the social norm will have no effect on the new value produced during a 10-hour workday. Such a change would instead act more like an increase in labor productivity in that more means of production will be transformed into final products and prices can be expected to fall.

Simple Labor versus Complex Labor

Throughout this entire discussion, it has been assumed that the labor that is being performed is of a very simple variety. That is, no special skill or training is required to perform this specific type of labor, which we will call **simple labor**. Of course, most types of labor require at least some basic training and many types of labor require years of prior education and training if they are to be performed well. These more skilled types of labor we will refer to as **complex labor**.

The existence of complex labor appears to create a difficulty for Marxian economics. If one hour of simple labor (e.g., sweeping floors) creates the same amount of value as one hour of complex labor (e.g., surgical labor), then this theory appears to be flawed. Recall, however, that SNALT is not the same as clock time, and so it is possible that one hour of surgical labor might create 100 times as much value as one hour of unskilled labor.

To understand how Marxian value theory can address these issues, let's consider a numerical example. Suppose that a person goes to a technical school for four years and learns to produce a specialized commodity. The number of hours spent in school during these four years might be 8,320 hours, which may be calculated as follows:

$$total\ hours\ of\ education = (4\ years)(\frac{52\ weeks}{year})(\frac{5\ days}{week})(\frac{8\ hours}{day}) = 8,320\ hours$$

Suppose the worker then works for 40 years producing the specialized commodity. During this 40-year period, the number of hours worked may be calculated in a similar way:

$$total\ hours\ of\ work = (40\ years)(\frac{52\ weeks}{year})(\frac{5\ days}{week})(\frac{8\ hours}{day}) = 83,200\ hours$$

The total value created during the working life of this person may be expressed in SNALT as the sum of the hours spent in training plus the hours spent working. This calculation is as follows:

$$Total\ value\ created = 8,320\ hours + 83,200\ hours = 91,520\ hours$$

Further suppose that the worker produces 9,152 use values during her entire working life. To keep the example simple, let's ignore the value of the means of production by assuming that the constant capital advanced is equal to zero. We can use this information to calculate the value (or price) per unit of the commodity produced in terms of SNALT as follows:

$$Price\ (in\ SNALT) = \frac{91{,}520\ hours}{9{,}152\ units} = 10\ hours\ per\ unit$$

If we assume a MELT of \$6 per hour, then the price of the commodity will be \$60 per unit (=\$6/hour times 10 hours/use value), and the total value of the worker's lifetime product will be \$549,120 (=\$60/unit times 9,152 use values).

If we next consider an unskilled worker who works for 40 years performing simple labor and producing a similar, albeit unspecialized commodity, then we can see what contribution the first worker's training makes to the production of value. Let's assume that the unskilled worker also produces 9,152 units of the unspecialized commodity. Since the worker works for 40 years, she has performed 83,200 hours of work, just like the skilled worker. The value of each unit of the unspecialized commodity may be calculated in terms of SNALT as follows:

$$Price\ (in\ SNALT) = \frac{83{,}200\ hours}{9{,}152\ units} \approx 9.09\ hours\ per\ unit$$

Using the same MELT of $6 per hour, the price of the unspecialized commodity will be about $54.54 (=$6/hour times 9.09 hours per use value), and the total value of the worker's lifetime product will be about $499,150 (≈ $54.54/unit times 9,152 use values), ignoring some rounding error here.

This example shows rather clearly that the skilled worker produces a more valuable product than the unskilled worker. The difference in the value created occurs because the skilled worker creates more value in the same 40-year period. This enhanced value-creating potential is not the result of a more intense labor process or a longer working day. The superior ability of the skilled worker to create value exists because the hours the worker has spent acquiring specialized knowledge are labor hours that were necessary for the worker to produce and reproduce her labor-power. Just as work is required to produce the means of subsistence the worker needs to perform labor each day, work is also required to produce the knowledge that the worker uses to produce commodities each day. In summary, the value-creating potential of complex labor increases with the educational requirements of the specialized labor process that requires that special type of labor.[4]

Following the Economic News[5]

Samsung is the world's largest producer of smart phones. According to a New York-based watchdog group called

China Labor Watch, underage workers have been employed at one of Samsung's suppliers in Dongguan, China. The Shinyang factory is accused of hiring children and underage students during the busiest times when it is in "urgent need of labor." The executive director of China Labor Watch claims that the underage workers are only 15 years old and have been using the identification of other people. According to neoclassical economic theory, this factory is a wage-taker and would be run out of business if it paid a wage higher than the market wage. If child labor laws were strictly enforced, however, then it would be possible to prevent the market wage from falling to a very low level. Marxian theory also sheds light on this situation. In a written statement by China Labor Watch, the underage workers are usually employed for three to six months and are only paid for 10 hours a day even though they work for 11 hours. This case is a perfect example of absolute surplus value production. This example is rather unusual, however, because typically it appears that workers are being paid for the entire value produced during the workday. For example, a worker might receive an hourly wage of $5 per hour for 11 hours. However, the worker might produce $2 of surplus value during each hour as well. In that case, the variable capital advanced for the day is $55 and the surplus value produced is $22. The worker has only been paid for part of the workday, but it appears that the worker is paid for all work performed. In the case of the Shinyang factory, it does not even *appear* that workers are paid for the entire value that they produce during the workday. For the reasons already mentioned, the surplus labor performed almost certainly exceeds the one hour of transparent surplus labor time.

Summary of Key Points

1. For a perfectly competitive employer, the total resource cost (TRC) curve grows continuously with employment, but the average resource cost (ARC) and the marginal resource cost (MRC) curves are identical to the labor supply curve facing the firm due to the constant wage rate.
2. The marginal revenue product (MRP) curve of a perfectly competitive employer may be calculated by multiplying the product price by the marginal product of labor.
3. The profit-maximizing rules (MRP = MRC and only operate when $w \leq$ ARP) leads to the conclusion that the MRP curve below the maximum ARP is the perfectly competitive employer's labor demand curve.
4. A shift of the labor demand curve may result from either a change in the product price or a change in the marginal product of labor.
5. The individual worker's labor supply curve is determined by utility maximization as the worker considers the tradeoff between income and leisure while faced with a time constraint.
6. When a backward bending labor market supply curve exists, both a stable equilibrium and an unstable equilibrium may exist.
7. In a monopsony labor market, the MRC rises more quickly than the ARC because the monopsonist must pay all workers a higher wage when an additional worker is hired.
8. In a monopsony labor market, economic exploitation exists because the MRP exceeds the wage paid, but in a bilateral monopoly labor market, the degree of exploitation depends on

the relative strength of the employer and the union.

9. In Marxian theory, an increase in productivity in industries that produce workers' means of subsistence lowers the prices of those commodities and increases the production of relative surplus value, but when the productivity increase occurs in other industries, it only causes a reduction in commodity prices.
10. In Marxian theory, an increase in the length of the working day increases the amount of absolute surplus value produced, but it leaves commodity prices unchanged.
11. In Marxian theory, an increase in the intensity of labor increases the amount of surplus value produced during a working day of a given length, but leaves commodity prices unchanged.
12. In Marxian theory, complex labor creates a larger amount of value in a specific period than simple labor in the same amount of time.

List of Key Terms

Factor markets (input markets or resource markets)

Wage-taker

Wage elasticity of labor supply

Total resource cost (TRC)

Average resource cost (ARC)

Marginal resource cost (MRC)

Marginal revenue product (MRP)

Average revenue product (ARP)

Labor market demand curve

Derived demand

Time constraint

Income/leisure tradeoff

Substitution effect

Income effect

Stable equilibrium

Unstable equilibrium

Marginal productivity theory of income distribution

Monopsony

Industrial union

Craft union

Bilateral monopoly

Simple labor

Complex labor

Problems for Review

1.Complete the missing information for the perfectly competitive employer represented in the table below. Assume the product price is $2 per unit. Then determine the profit-maximizing employment level.

Problems for Review: Problem 1

L	w	TRC	MRC	ARC	MP	MRP
0	8				--	
1	8				15	
2	8				18	
3	8				17	
4	8				14	
5	8				8	
6	8				4	

2. Suppose T = 20 and w = \$3 per unit of labor. Derive the equation of the time constraint beginning with the fact that T = h+l and Y = wh. When utility is maximized, what will the slope of the indifference curve be that is just tangent to the time line?

3. Complete the missing information for the monopsony employer represented in the table below. Assume the product price is \$2 per unit. Then determine the profit-maximizing employment level and wage rate.

Problems for Review: Problem 3

L	w	TRC	MRC	ARC	MP	MRP
0	8				25	
1	10				29	
2	12				22	
3	14				18	
4	16				11	
5	18				1	
6	20				-6	

4. Suppose the working day is 11 hours, the variable capital is \$32, the constant capital is \$124, and the MELT is \$4 per hour of SNALT. Also, assume that 50 pounds of the product are produced in one day, and this sector does not produce wage commodities.

- What is the current price per pound of the product?
- Suppose labor productivity rises in the wage commodities sector causing the variable capital to fall to \$24. What will happen to the surplus value, the necessary labor, and the product price as a result?
- Returning to the original conditions, suppose that a 20% increase in labor productivity occurs in this industry alone. What will happen to the product price, the surplus value, and the necessary labor in this case? Be sure to account for the change in the amount of use values produced and the change in the constant capital advanced.
- Returning to the original conditions, suppose that the working day is extended from 11 hours to 12 hours. What percentage increase in the length of the workday is this change? What will happen to the surplus value, the constant capital, and the product price as a result?
- Returning to the original conditions, suppose that the intensity of labor increases by 10%. This change is equivalent to how much of a change in the length of the workday? What is the new surplus value, the new constant capital, and the price of the commodity?

5. Suppose a worker spends 2 years in technical school. The training involves a 7-hour workday for 6 days each week during the 52 weeks in the year. The worker then works 8 hours per day and 7 days per week for 30 years. If the constant capital advanced during those 30 years equals $200,000 and the MELT is $9 per hour, then what is the total value produced? Also, if 80,000 use values are produced during the 30 years, then what is the value (price) of the product?

Notes

1. Prof. David Ruccio's presentation of the neoclassical theory of labor supply in his introductory economics class at the University of Notre Dame in the early 2000s inspired the presentation in this section. I served as Prof. Ruccio's teaching assistant at the time.
2. Chiang and Stone (2014), p. 305-306, represent an exception to the usual rule. They refer to the "monopsonistic exploitation of labor" and even include a box devoted to Marx's critique of capitalism. They do not emphasize, however, that Marx's condemnation of capitalism applies equally to intensely competitive market conditions. They refer to the term "exploitation" as loaded, which seems to imply that it should be used with caution. The caveat is not surprising. The authors are one step away from entering a competing discourse that neoclassical economists generally refuse to acknowledge.
3. For an excellent account of the Homestead strike, see Wolff, Leon (1965).
4. In this example, we have ignored the labor embodied in school supplies and equipment. The intensity of schooling is another

difficult aspect of the problem, but it would need to be considered as well.

5. Lee, Min-Jeong. "Samsung Faces New Child-Labor Allegations." *The Wall Street Journal*. July 10, 2014.

PART IV

PART THREE: PRINCIPLES OF MACROECONOMIC THEORY

CHAPTER 12

MACROECONOMIC MEASUREMENT

Goals and Objectives:

In this chapter, we will do the following:

1. *Measure* the amount of poverty in an economy
2. *Explore* the way that income and wealth inequality are measured
3. *Analyze* two ways of measuring the aggregate output of an economy
4. *Examine* two critiques of national income accounting
5. *Define* the labor force and the unemployment rate
6. *Investigate* the two primary methods of measuring the aggregate price level
7. *Explain* the meaning of the inflation rate
8. *Inspect* historical movements of the key macroeconomic variables over time

In Part II, we investigated many theories that are regarded as microeconomic theories because they concentrate on individual consumers, workers, savers, and business enterprises. In Part III, we turn our attention to macroeconomic theories that concentrate on much broader changes in the economy, including changes in the behavior of households, governments, industries, foreign nations, and social classes. These theories use different economic variables than microeconomic theories because the subject matter is so much broader. To understand these theories then, it is necessary first to discuss how macroeconomic variables are measured. This chapter thus concentrates entirely on the issue of macroeconomic measurement and will set the stage for all the theories that we explore in Part III. The chapter discusses how to measure poverty, income inequality, wealth inequality, aggregate output, the labor force, the unemployment rate, the aggregate price level, and the rate of inflation. After each macroeconomic variable is defined and the method of its measurement is described, its historical pattern is considered. The historical observations will also point us in the direction of interesting questions that can only be answered with the help of the theoretical frameworks that are developed in later chapters. Also in this chapter, we will consider two important critiques of national income accounting, which is important because it shows that disagreements within economics are not confined to the realm of theory but also arise around questions of measurement.

The Measurement of Poverty

The well-being of a nation depends on many factors. Neoclassical economists argue that people have unlimited wants. They do not draw a clear distinction between wants and needs. The lack of this distinction in neoclassical theory is one source of disagreement between neoclassical and heterodox economists.

Heterodox economists sometimes argue that basic needs for food, clothing, medical care, and housing are fundamentally different from preferences for fine clothes, jewelry, and expensive works of art. It is not simply the *strength* of the preference, according to this heterodox view, but the *nature* of the preference that separates needs from wants.

Because this textbook takes heterodox approaches seriously, it will approach the subject of macroeconomic measurement in a way that sharply deviates from most neoclassical economics textbooks. Neoclassical economics textbooks generally begin the discussion of macroeconomic measurement with an explanation of how the total output of a nation is measured. Goods and services of all types are lumped together according to their market values and no effort is made to distinguish between goods and services that fulfill basic human needs and the goods and services that are desirable but not essential for human life. To take the heterodox perspective seriously then, this chapter acknowledges a distinction between basic needs and inessential wants. It does so by starting with poverty measurement as a measure of the well-being of a nation. That is, the welfare of a nation's people is evaluated according to how well the population meets its basic needs.

The U.S. Census Bureau is the government body responsible for the measurement of poverty in the United States. It uses an official poverty measure and a supplementary poverty measure and each is based on "estimates of the level of income needed to cover basic needs."[1] To calculate the official poverty rate, the U.S. Census Bureau calculated the amount of money that a household spent on food in 1963 and then tripled it while adjusting it for inflation in later years and for differences in family size, family composition, and age of the householder.[2] This amount of money income is called the **poverty threshold**. According to the U.S. Census Bureau, 48 different poverty thresholds exist because families are so different according to size and age.[3] In any case, the measure suggests that a household needs to spend a full 1/3 of its income on food, leaving 2/3 for all other expenses.

Once the poverty threshold is known, it is possible to determine whether a family lives in poverty. The U.S. Census Bureau calculates the **Ratio of Income to Poverty** by dividing total family income by the poverty threshold as follows:[4]

$$Ratio\ of\ Income\ to\ Poverty = \frac{Total\ Family\ Income}{Poverty\ Threshold}$$

The following definitions are used:

$$Ratio\ of\ Income\ to\ Poverty < 1 \Rightarrow poverty$$
$$1 \leq Ratio\ of\ Income\ to\ Poverty \leq 1.24 \Rightarrow near\ poverty$$
$$Ratio\ of\ Income\ to\ Poverty \leq 0.50 \Rightarrow deep\ poverty$$

In words, if the ratio of income to poverty is less than one, then the family is living in **poverty** because its income is below the poverty threshold. If the ratio of income to poverty is greater than or equal to one but less than 1.24, then the family is living at a **near poverty** level because its income has not reached 125% of the poverty threshold. Finally, if the ratio of income to poverty is less than or equal to half of the poverty threshold, then the family is living in **deep poverty**.[5]

The U.S. Census Bureau also provides a helpful example to illustrate the calculation.[6] A similar example is provided below:

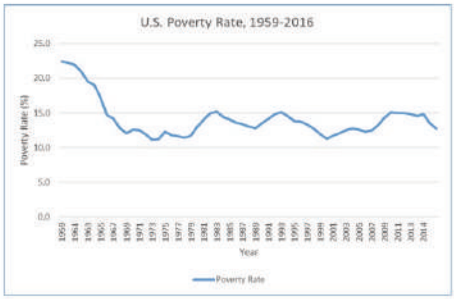
Suppose that a family of five earns $28,000 per year. The 2016 poverty threshold for a family of five was $29,360. The ratio of income to poverty in this case is $28,000/$29,360 = 0.9537. Because the ratio of income to poverty is less than one but greater than 0.50, the family is living in poverty although not in deep poverty. The U.S. Census Bureau also defines the **income deficit** (if negative) or the **income surplus** (if positive) as the difference between family income and the poverty threshold as follows:[7]

$$Income\ deficit\ (or\ surplus) = Income - Threshold = 28,000 - 29,360 = -1,360$$

In other words, the family of five would require $1,360 to meet the threshold and move from poverty to near poverty.

Finally, the **official poverty rate** refers to the percentage of the population that lives below the poverty threshold. Over time, the U.S. official poverty rate has fluctuated as shown in Figure 12.1.

Figure 12.1: The U.S. Official Poverty Rate, 1959-2016



Source: Source: U.S. Bureau of the Census. "Table 2: Poverty Status of People by Family Relationship, Race, and Hispanic Origin: 1959 to 2016." Web. Accessed on April 14, 2018. https://www.census.gov/data/tables/time-series/demo/income-poverty/historical-poverty-people.html

As Figure 12.1 shows, the U.S. poverty rate fell significantly during the economic expansion of the 1960s but rose during the recessions in the early 1980s and early 1990s. It also declined during the economic expansion of the 1990s but rose again after the 2001 recession and even more during the Great Recession. The poverty rate thus seems to follow a somewhat **countercyclical** movement, which means that it rises during recessions and falls during expansions.

The official poverty rate has been in use for a half century, but it has some serious shortcomings. The Institute for Research on Poverty at the University of Wisconsin-Madison has summarized the most common criticisms of the official poverty measure, a few of which are listed below:[8]

1. It only represents a headcount, but it does not measure "the depth of economic need."

2. It omits taxes and medical expenses and does not include noncash income like food assistance.
3. It does not account for geographic differences in the cost of living throughout the U.S.

We can also add to this list the omission of many people such as those in prison or nursing homes, homeless people, and foster children under age 15.[9]

Because of the problems with the official poverty measure, by 2008, New York City and other cities were developing their own poverty measures.[10] The official poverty rate has become increasingly irrelevant because as Rebecca Blank explains, food prices have fallen significantly and housing and energy prices have risen.[11] The poverty threshold has become less meaningful as a result. Resolving these issues is of great importance because food stamp eligibility depends on it, and some federal block grants to states depend on state poverty rates.[12]

To address these issues, the U.S. Census Bureau introduced a supplemental poverty measure in 2011. The **supplemental poverty measure** offers "a more complex statistical understanding of poverty by including money income from all sources, including government programs, and an estimate of real household expenditures."[13] The supplemental poverty measure is also linked to poverty thresholds but the thresholds tend to be higher than the official poverty thresholds.[14] The new measure has other benefits, such as its ability to demonstrate the impact of specific safety net programs on poverty rates.[15] Nevertheless, as its name suggests, the supplemental poverty measure has not yet replaced the official poverty measure. Instead, it continues to be used as an additional tool for the measurement of poverty.

The Measurement of Income Inequality and Wealth Inequality

In neoclassical economic theory, a person's well-being is asserted to depend only on his own consumption level with greater levels of consumption representing greater amounts of satisfaction or utility. Heterodox economists often criticize this way of thinking because it ignores the impact that unequal consumption levels may have on human well-being. This section is also committed to taking the heterodox perspective seriously and so will consider the two measures of well-being that are most relevant in this connection: measures of income inequality and wealth inequality.

The amount of inequality that exists in society directly affects human well-being. Those with lower incomes or less wealth experience envy and feel dissatisfied with what they have. Those feelings arise because others have more and those with more often enjoy putting it on display for others to see. Those with lower incomes or less wealth may devote a great deal of time and effort trying to acquire more. They may turn to illegal activities such as illegal drug sales or burglary to accumulate more and overcome such feelings. Depression and anxiety may also be a result of slipping behind others in the race to accumulate material possessions. To overcome these feelings, many people turn to shortcuts such as gambling and playing the lottery. Because such solutions rarely lead to lasting gains for people, the pressure to find a solution becomes that much greater.

On the other hand, those with high incomes or great wealth become the subjects of envy and are placed in a defensive position. They must devote effort to justifying their high incomes or great wealth. Economic theory may serve this end insofar as it provides theoretical explanations for the incomes and wealth levels that emerge in capitalist societies. Nevertheless, many with great incomes and wealth will put it on display so that it becomes an object of envy for others. Such displays are what Thorstein Veblen called **conspicuous consumption** and might include expensive artwork, mansions, boats, sportscars, jewelry, and vacations. Others with great income and wealth separate themselves from the rest of the population in gated communities or high-rise apartments.

At all levels, the preoccupation with having more leads people to forget about other aspects of life such as family relationships, which often suffer because of the focus on material gain. The beauty of nature and the joy of hobbies are also forgotten as people seek ways to accumulate more wealth and to elevate themselves above their peers. Great wealth can also lead to the exploitation of labor-power from a Marxian perspective as capital is put in motion to produce surplus value. Because income inequality and wealth inequality are so important to our economic well-being, it makes sense to explore the primary method of measuring them.

One method of measuring income inequality is to use a statistic called the quintile ratio. The **quintile ratio** is the ratio of the income of the top fifth of the population to the income of the bottom fifth of the population. The ratio ignores the middle 3/5 of the population, but it helps us to see just how much of a spread exists between the top income earners and the bottom income earners. The higher the quintile ratio, the higher is the degree of income inequality. For example, a quintile ratio of 5 implies that the top income earners have five times the income of the bottom income earners. If the quintile ratio rises to 6, then the top income earners have six times the income of the bottom income earners, and inequality has increased.

Table 12.1 shows the quintile ratios for several countries in 2018.

Table 12.1: Income Inequality Quintile Ratios in 2016

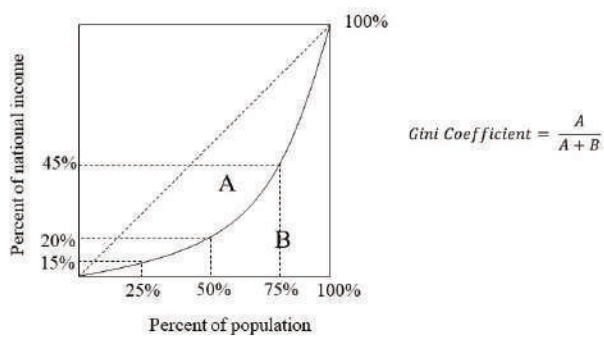
Country	Quintile Ratio
Colombia	17.3
Mexico	10.8
United States	9.1
Russia	8.2
Greece	7.6
Canada	5.8
Japan	5.4
India	5.3
Germany	4.6
Norway	3.8

Note: The quintile ratio is calculated as the ratio of the mean income of the richest 20% to the mean income of the poorest 20% of the population.
Source: United Nations Development Programme. Human Development Reports. Income inequality, quintile ratio. Web. Accessed on April 14, 2018. http://hdr.undp.org/en/indicators/135106

Table 12.1 arranges the countries from the least equal to the most equal as reflected in the falling quintile ratios as you move down the column. In Table 12.1, Colombia has the highest degree of income inequality, and Norway has the least income inequality of the countries in the table.

A second method of measuring income inequality is with something called the **Lorenz Curve**. The Lorenz Curve at any point shows the percentage of national income that a specific percentage of the population holds. It thus provides a graphical representation of the distribution of income. Figure 12.2 shows an example of a Lorenz Curve.

Figure 12.2: The Lorenz Curve



In Figure 12.2, the first 25% of the population holds 15% of the income. The first 50% of the population holds 20% of the income. The first 75% of the population holds 45% of the income. Finally, 100% of the population holds 100% of the income.

The 45-degree line has a special role to play relative to the Lorenz Curve. The 45-degree line represents perfect equality. It shows that 25% of the population holds 25% of the income. 50% of the population holds 50% of the income. 75% of the population holds 75% of the income. 100% of the population holds 100% of the income.

Therefore, the further away from the 45-degree line the Lorenz Curve is, the more income inequality is implied.

It is possible to measure the extent of the deviation of the Lorenz Curve from the 45-degree line. Two areas have been marked in the graph: Area A and Area B. When Area A is larger and Area B is smaller, the Lorenz Curve is further from the 45-degree line, and more income inequality exists. When Area A is smaller and Area B is bigger, then the Lorenz Curve is closer to the 45-degree line and less income inequality exists. To measure the extent to which the Lorenz Curve deviates from the 45-degree line, economists use something called the Gini Coefficient. The **Gini Coefficient** is calculated as Area A divided by the sum of Areas A and B:

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

The extreme values of the Gini Coefficient are zero and one. When the Lorenz Curve coincides with the 45-degree line, Area A is equal to zero and so the Gini Coefficient is equal to zero, which indicates **perfect income equality**. When the Lorenz Curve perfectly coincides with the lower right angle, Area B is equal to zero and so the Gini Coefficient is equal to 1, which indicates perfect income inequality. **Perfect income inequality** means that one person has all the income and the rest of the population has zero income. In general, the Gini Coefficient will fall somewhere in between these extremes and is usually between 0.20 and 0.50. Extreme cases are a bit higher or lower.

Table 12.2 shows estimates of the Gini Coefficient for several years for the United States.

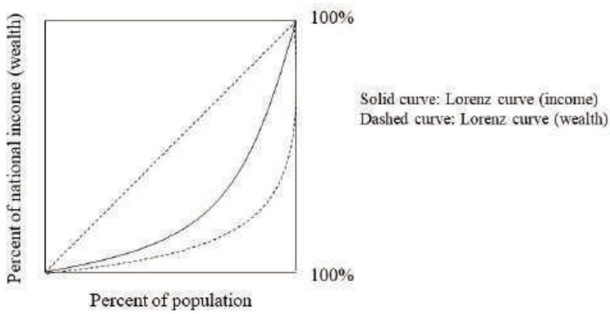
Table 12.2: U.S. Gini Index, 1979-2013 (not seasonally adjusted)

Year	Gini Index (annual)
1979	34.6
1986	37.5
1991	38.2
1994	40.2
1997	40.8
2000	40.4
2004	40.5
2007	41.1
2010	40.4
2013	41.0

Note: The Gini Index is equal to the Gini coefficient multiplied by 100 where an index value of 100 represents perfect inequality. Source: World Bank, GINI Index for the United States [SIPOVGINIUSA], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/SIPOVGINIUSA, April 13, 2018.

Table 12.2 shows clearly that the level of income inequality in the U.S. has worsened over time. It is also possible to create a Lorenz Curve to represent the distribution of wealth and a Gini Coefficient to measure the extent of wealth inequality. In the U.S., the distribution of wealth has been much more unequal than the distribution of income. Figure 12.3 places Lorenz Curves representing the distribution of income and the distribution of wealth on the same graph.

Figure 12.3: Lorenz Curves Representing Income Distribution and Wealth Distribution



Because the wealth distribution Lorenz Curve is further from the 45-degree line than the income distribution Lorenz Curve, we can infer that the distribution of wealth is more unequal than the distribution of income. This inequality in the distribution of wealth is only expected to worsen. According to a new analysis that the House of Commons Library conducted, if the current pattern continues, then the top 1% of the global population will own 64% of global wealth by 2030.[16]

The Measurement of Aggregate Output

We now turn to the primary measure of macroeconomic performance among neoclassical economists, which is a measure of the aggregate output of the economy that is called **Gross Domestic Product (GDP)**. GDP is intended to give us a sense of the size of the economic pie of a nation. It is one of the major components of the **National Income and Product Accounts (NIPA)**. The

U.S. Bureau of Economic Analysis (BEA) within the U.S. Department of Commerce maintains the NIPA and publishes quarterly estimates of U.S. GDP.

To be precise, GDP represents the total market value of all final goods and services produced in a year within the national boundaries of a nation. Final goods and services refer to goods and services that are sold for final consumption. It should also be noted that GDP is a **flow variable** because it is measured per period such as a year. Figure 12.4 shows a **production possibilities frontier (PPF)** for a simple economy with just two final goods: apples and oranges.

Figure 12.4: Two Combinations of Goods on a Production Possibilities Frontier (PPF) in 2001 and 2002

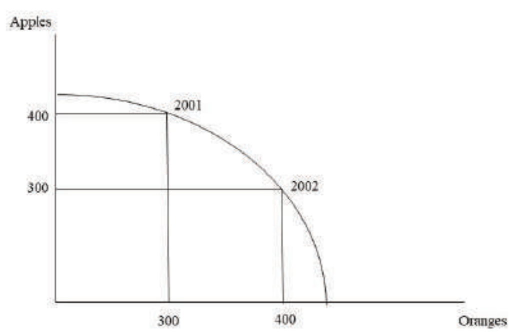


Figure 12.4 shows two combinations of apples and oranges that the economy produces in two different years. In 2001, it produces 400 apples and 300 oranges. In 2002, it produces 300 apples and 400 oranges. The reader should recall that the quantities of each good are

measured in real terms (i.e., so many apples and so many oranges). It is not meaningful to add up all the apples and oranges in a year because they are qualitatively different goods. Even if we are satisfied adding together different types of fruit, if the goods produced in this economy included apples and automobiles, then adding these goods together would really make no sense. In general, the fact that differences exist among the units in which each good is measured prevents us from adding together the real quantities.

Neoclassical economists resolve this problem through the assignment of weights to each good, which makes possible their conversion into a common unit and their aggregation. The natural weights to use are the market prices of the goods. More valuable goods, like automobiles, will be assigned greater weights and less valuable goods, like apples, will be assigned smaller weights. Table 12.3 adds price information to our example of an economy that produces apples and oranges.

Table 12.3: Gross Domestic Product (GDP): An Example

Year	Physical Output (Apples)	Physical Output (Oranges)	Price of Apples	Price of Oranges	GDP
2001	400	300	$1.00	$1.25	$775
2002	300	400	$1.00	$1.25	$800

If we multiply each real quantity of a good by the market price of the good, then we obtain a dollar value of that good for the year. We then add the dollar values of apples and oranges for that year and we obtain the aggregate output or GDP for this simple economy. This measure of aggregate output makes it possible for us to compare the size of the economic pie across two years. Since GDP has risen from $775 to $800, we conclude that GDP has risen from 2001 to 2002. Without the common metric that money provides, it would not be possible to draw any conclusions about the change in aggregate output between 2001 and 2002.

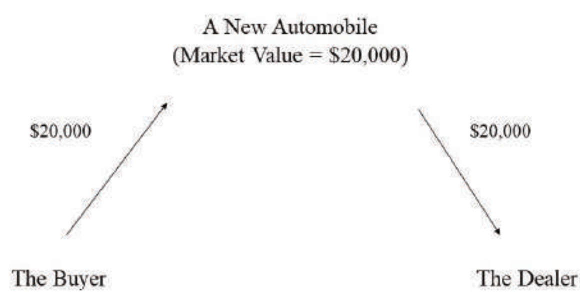
It is important to ask why economists limit the measurement of aggregate output to final goods and services. The values of **intermediate goods**, or goods that become part of other goods during production, are specifically excluded from the GDP calculation. The reason for the exclusion of the values of intermediate goods is that their inclusion would lead to a problem referred to as **double counting**. For example, suppose that a tire manufacturer purchases rubber from a supplier at a price of $150. The tire is then manufactured and sold to an automobile manufacturer for $250 who uses it to produce an automobile. The automobile is then sold to a consumer for $20,000. The automobile is the final good in this scenario, and the rubber and tire are intermediate goods. Therefore, only the value of the automobile is counted as part of GDP and the values of the rubber and the tire are intentionally omitted from the calculation. Why? The reason is that the $20,000 price of the automobile includes the value of the tire, which includes the value of the rubber. The supplier has added $150 to the value of the rubber through its production process (assuming it is the first stage of production). The tire manufacturer then adds another $100 to the value of the rubber, which results in a tire worth $250. The automobile manufacturer then adds additional value to the tire because it is now a part of a finished automobile. Let's suppose that $400 is the value of the tire, which makes up part of the $20,000 sale price of the automobile. The automobile manufacturer has thus added another $150 of value to the tire. The reason for excluding the values of the rubber and the tire should be clear. The $400 tire, which is part of the sale price of the automobile, already includes the value of the rubber sold to the tire manufacturer and the value of the tire sold to the automobile manufacturer. If we count the value of the rubber and the value of the tire in the calculation of GDP, then we will be counting the value of the rubber three times and the value of the tire two times! To avoid multiple counting, we only add the value of the final good or service when calculating GDP. An alternative method is to add up the values added at each stage of production. In this case, we would add $150 for the value of the rubber, $100 for the value that the tire manufacturer adds, and $150 for the value that the automobile manufacturer adds to the tire due to the production of the finished automobile. Of course, we would then need to add the remaining $19,600 of value that the auto manufacturer adds with labor and other component parts to obtain the $20,000 contribution to GDP. Either of the two methods avoids double counting, and arguably results in a better approximation of the contribution of these goods to the national economic pie.

In addition to the exclusion of the values of intermediate goods, several other exclusions apply to the calculation of GDP. National income accountants exclude government transfer payments. **Government transfer payments** include social security benefits and public assistance of all kinds. Because they do not represent a payment for a real good or service, they do not count in the calculation of GDP. National income accountants also exclude private transfer payments from the calculation of GDP. **Private transfer payments** include monetary gifts and charitable donations. When a donor makes a charitable donation, she typically receives a letter from the charity thanking her for the contribution. The letter also states that the organization did not grant any goods or services in exchange for the donation. Because the donation does not reflect any current production of goods or services, GDP should not include it. National income accountants also exclude sales and purchases of financial assets, such as stocks and bonds. It is possible to link stocks and bonds to production processes, but these assets only represent claims to the assets of a corporation and so are not included in GDP.

Finally, used goods are also excluded from the GDP calculation because the current year GDP or the previous year's GDP includes the values of the goods when sellers sold them the first time. For example, the sale of a 2014 Ford Escape in 2018 should not be included in the GDP for 2018 because the 2014 GDP already included it. The sale of used goods represents the redistribution of existing output rather than the production of new output. For that reason, the GDP calculation excludes the values of used goods.

From a conceptual perspective, two methods exist for thinking about the measurement of aggregate output. The two methods stem from what is an identity in neoclassical economics, namely that income and expenditure are always equal as shown in Figure 12.5.

Figure 12.5: Income and Expenditure: A Neoclassical Identity



If I purchase that new 2014 Ford Escape for $20,000 in 2014, then we can think about my expenditure of

$20,000, which is equal to the value of that final good. We can also think about it from the perspective of the dealer who receives an income of $20,000 when she sells the automobile. At a macroeconomic level, whether we add up all the expenditures on final goods or whether we add up all the incomes received from the sale of final goods, we should obtain the same measure of aggregate output. This result must hold true because of the **income-expenditure identity**.

To delve deeper into the **expenditures approach** to the measurement of GDP, national income accountants divide aggregate expenditures into four major categories: personal consumption expenditures (C), gross private domestic investment (I), government purchases (G), and net exports (X_n). If we add together these four values, we obtain GDP as follows:

$$C + I + G + X_n = GDP$$

Personal consumption expenditures include expenditures on durable goods, nondurable goods, and services. The consumption of **durable goods** typically requires more than three years, and includes such items as automobiles and household appliances. The consumption of **nondurable goods** typically requires less than three years, and includes such items as food, clothes, and fuel. Finally, **services** include nontangible commodities like legal services, medical services, and childcare.

Gross private domestic investment includes several types of expenditure as well. Expenditures on **final capital goods** that businesses incur are included in this category. When a business purchases a machine, for example, it is considered a final capital good because the machine does not become physically incorporated into another product. Its use in the production process is its final use. One potential complication here is that the value of a final capital good is gradually transferred to the value of the good that it is used to produce. As the machine depreciates, that value must pass to the value of the final product because it represents a cost of production. Later in this section, we will consider how national income accountants address this issue.

Residential fixed investment is another type of investment expenditure in the national income accounts. It refers to all expenditures incurred in the purchase of newly constructed homes. When homes are resold, they are not included in the GDP calculation because that would represent double counting. Previous home construction was already counted once when the homes were sold for the first time. The reader might find it odd that homes are considered an investment expenditure rather than a consumption expenditure. The reason is that investment expenditures are a positive contribution to the nation's stock of capital and houses may be thought of as capital goods. In neoclassical theory, capital goods are goods used to produce other goods. In the case of housing, houses produce a flow of services over time. That is, a home creates a space for a person to live that can benefit that person for many years. The house thus contributes to the production of this service and so the house may be thought of as a capital good. **Business fixed investment** is another category of investment expenditure in the national income accounts. It refers to expenditures incurred in the construction of new factories, production plants, and office buildings. Business investments of this kind also make possible the production of other goods and services and so represent an increase in the nation's capital stock.

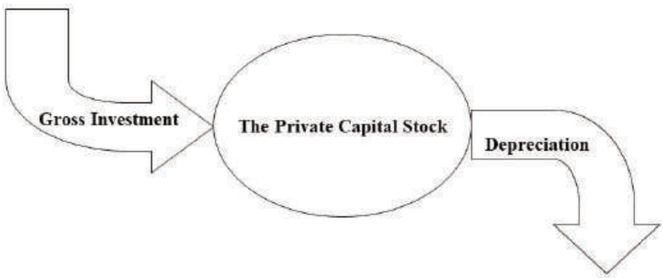
Inventory investment is another type of investment expenditure in the national income accounts. It is calculated as changes in inventories in a year. Business inventories expand when firms have unsold goods at the end of a calendar year. They store these goods with the hope of selling the goods in the next year. Even though these goods are not sold to the public, they do represent new production of final goods and should be counted in GDP. Hence, national income accountants include new additions to inventories when calculating GDP. It is as if the businesses purchase the goods even though no money changes hands. On the other hand, some businesses will sell goods in the current year that were produced in a previous year and became part of their business inventories in that previous year. Because these goods were already counted as part of a previous year's GDP since they represented additions to inventories at that time, these sales should be subtracted in the calculation of GDP. One might wonder why they need to be subtracted as opposed to simply ignored. They must be subtracted because when personal consumption expenditures are calculated, they include all goods and services sold to consumers, which might include goods that were produced in a previous year and became part of business inventories at that time. The subtraction at this stage allows national income accountants to remove them from the calculation of GDP. Inventory investment is thus calculated as follows:

$$Inventory\ Investment = New\ additions\ to\ inventories - reductions\ in\ inventories$$

It is possible for inventory investment to be positive if new additions to inventories outweigh reductions in inventories in a year. It might also be equal to zero if the additions and reductions perfectly balance. Finally, it might be negative, if reductions in inventories are so large that they exceed new additions to inventories. In the last case, negative inventory investment will cause GDP to be smaller.

At this point it might be helpful to consider how the nation's stock of private capital changes over time.[17] The nation's **private capital stock** refers to all the privately-owned machinery, homes, factories, office buildings, production plants, apartment buildings, etc. at a specific point in time. Gross private domestic investment causes the private capital stock to grow. At the same time, depreciation causes the private capital stock to contract. **Depreciation** refers to the gradual wearing out of capital over time due to use or lack of use. The relationship between gross investment and depreciation in a year determines the net impact on the private capital stock. Gross investment may be thought of as an inflow and depreciation may be thought of as an outflow relative to the private capital stock. Figure 12.6 shows these relationships.

Figure 12.6: The Effect of Gross Investment and Depreciation on the Private Capital Stock



Since gross investment and depreciation cause annual changes in the size of the private capital stock, they are flow variables. The private capital stock is obviously a **stock variable** because it is measured as of a point in time. If gross investment exceeds depreciation, then the private capital stock expands. That is, more is added to the private capital stock than is depleted. If gross investment is below depreciation, then the private capital stock contracts. That is, more of the private capital stock is wearing out than is being replaced. If gross investment and depreciation are equal, then the private capital stock remains constant for that year.

Returning to the components of aggregate expenditure that are used to measure GDP, the third type is **government purchases** of final goods and services. Federal, state, and local governments purchase consumer goods and services such as office supplies and computers for use in government buildings. Governments also make investment expenditures when they build new roads, bridges, schools, and office buildings. Because all these purchases represent purchases of final goods and services, we should include them in our calculation of GDP.

The final component of aggregate spending that is used to calculate GDP is net exports. **Net exports** are the difference between exports and imports (X – M). It is also referred to as the balance of trade or simply as the **trade balance**. It should be obvious why we add exports in the calculation of GDP. GDP is supposed to include the values of all domestically produced final goods and services in a year. When final goods and services are produced and exported, the expenditure that foreign buyers incur should be included in GDP. Imports of final goods and services, on the other hand, are produced outside the territorial boundaries of the nation and so should be included in the GDPs of foreign nations. The reader might wonder why we subtract imports in the calculation of GDP rather than simply ignoring them altogether. As with spending on goods produced in previous years, personal consumption expenditures might include spending on imported goods and services. No effort is made to exclude imported goods and services from that component of aggregate expenditure. Therefore, we subtract imports at this stage to ensure that they are not included in our GDP measure. Similarly, government purchases of final goods and services and business purchases of final capital goods might include purchases of imported goods. The subtraction of imports also allows us to exclude those values from our GDP calculation.

Because imports are subtracted in the calculation of net exports, it is possible for net exports to be negative. When a nation's imports exceed its exports, then we say that a **trade deficit** exists and net exports are negative. When a nation's exports exceed its imports, then we say that a **trade surplus** exists and net exports are positive. When the nation exports and imports the same amount, then we say that **balanced trade** exists and net exports are equal to zero.

Table 12.4 shows the numerical figures for each of the major components of aggregate expenditure in 2017.

Table 12.4: The Different Components of Aggregate Expenditure in the United States in 2017 (in billions of dollars)

	Amount (in billions of dollars)	Percentage of GDP
Personal Consumption Expenditures (C)	13,395.5	69.08%
Gross Private Domestic Investment (I)	3,212.8	16.57%
Government Purchases (G)	3,353.8	17.29%
Net Exports (X_n)	-571.6	-2.95%
Gross Domestic Product (GDP) in 2017	19,390.6	100%

Note: Some rounding error prevents the sum of the four components of expenditure from exactly equaling the total.

Source: U.S. Bureau of Economic Analysis. GDP and Personal Income: National Data. Table 1.1.5. Web. Last revised on March 28, 2018. Accessed on April 14, 2018. www.bea.gov.

The sum of personal consumption expenditures, gross private domestic investment, government purchases, and net exports is equal to GDP for that year. As Table 12.4 shows, GDP for 2017 was equal to $19.3906 trillion. In general, personal consumption expenditures tend to be about 2/3 (or about 66.67%) of GDP. This statistic is a frequently quoted statistic in the economic news of the nation. It is often stated that consumer spending makes up 2/3 of the economy. When people make this claim, they have in mind 2/3 of GDP. Government purchases are usually about 20% of GDP. Investment spending is typically about 10-15% of GDP. Net exports tend to be the smallest component of aggregate spending at about 3% and have been negative in recent decades. The percentages shown in Table 12.4 are approximately at these levels. Also, the reader should notice that the U.S. trade deficit is reflected in the negative value of net exports.

We next turn to the **income approach** to the measurement of GDP. The income approach adds up all the different flows of income that result from the sale of final goods and services. The largest income flow is **compensation** for American employees (i.e., wages and salaries). That is, when goods and services are sold, part of the revenue is used to pay employees of businesses. **Rental income** for American landlords is another major income category. Part of the revenue from the sale of final goods and services goes to pay rent for properties that are used in production. **Interest income** for American moneylenders is a third major income category. When final goods and services are sold, part of the revenue must be used to pay interest on loans. Finally, **profit income** for American businesses represents a major income flow. Part of the revenue from the sale of final goods and services businesses appropriate as profits. A portion of the profit income is for **unincorporated businesses** like **sole proprietorships** (one owner) and **partnerships** (multiple owners). Such businesses face **unlimited liability**. That is, if the business fails, then the owners' personal assets must be used to pay business debts. The rest of the profit income consists of **corporate profits** or the profits of incorporated business enterprises. Corporations issue and sell stock to the public. These business enterprises enjoy **limited liability**. If the firm fails, only the corporation's assets may be used to pay the firm's debts. The losses for the owners will be limited to the amount of money capital they contributed to the business. Corporate profits are subject to federal and state corporate income taxes and so a portion will be paid to the federal and state governments. Another portion may be distributed as **dividends** to the **shareholders**, who are the owners of the corporations. Finally, a third portion of corporate profits might be reinvested in the business and constitute what are called **retained earnings** because they are neither paid as taxes nor distributed to owners.

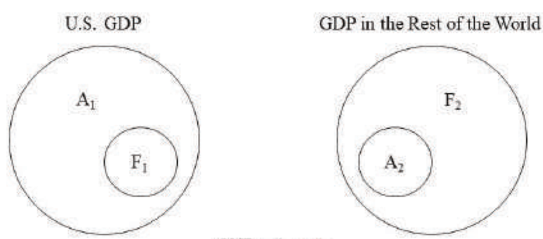
Three additional income flows must be considered before we arrive at our measure of GDP. When final goods and services are sold, a part of the revenue must be used to pay taxes on the sale. **Taxes on production and imports** include taxes, such as state sales taxes, excise taxes, and import tariffs. These income flows pass to federal and state governments. Another income flow is used to replace worn out capital. When businesses sell final goods and services, they set aside a portion of the revenue to repair and replace capital goods that have been used in production. A fund that represents the depreciation of the capital stock is thus another income flow that must be included in the GDP calculation.

The final income flow that must be included in the GDP calculation is a measure that national income accountants refer to as net foreign factor income. To understand this measure, we must recall that all compensation for employees, rental income, interest income, profit income and taxes on production and imports that were previously discussed flow to American citizens, businesses, and governments. **National income** is the sum of all these incomes flows as expressed below:

$$U.S.\ National\ Income = U.S.\ employee\ compensation + U.S.\ rental\ income + U.S.\ interest\ income + U.S.\ profit\ income + U.S.\ taxes\ on\ production\ and\ imports$$

That is, whether American workers and businesses are working and operating in the United States or in the rest of the world, their incomes are counted in these income measures. Because GDP measures all the income received within the territorial boundaries of the nation, to calculate U.S. GDP using the income approach, we must adjust aggregate income to account for the fact that some Americans are earning income abroad while some foreigners are earning income in the U.S. Figure 12.7 provides a diagram that shows why this adjustment must be made.

Figure 12.7: The Key Measures in the Transition from Gross National Product (GNP) to Gross Domestic Product (GDP)



$$GNP = A_1 + A_2$$

$$GDP = A_1 + F_1$$

$$GDP = GNP - (A_2 - F_1)$$

In Figure 12.7, the income that foreign citizens and businesses earn in the U.S. is denoted as F_1, and the income that American citizens and businesses earn outside the U.S. is denoted as A_2. Similarly, the income that American citizens and businesses earn in the U.S. is denoted as A_1, and the income that foreign citizens and businesses earn outside the U.S. is denoted as F_2. To calculate U.S. GDP, we need to subtract A_2 and add F_1 when starting with U.S. national income. This adjustment will allow us to calculate all the income earned within the geographical boundaries of the United States. We can now calculate **net foreign factor income** as follows:

$$Net\ Foreign\ Factor\ Income = Income\ of\ Americans\ working\ abroad\ (A_2) - Income\ of\ foreigners\ working\ in\ the\ U.S.\ (F_1)$$

To move us closer to the calculation of U.S. GDP, we need to *subtract* net foreign factor income from U.S. national income. Subtracting net foreign factor income will remove the income of Americans working abroad and add the income of foreigners working in the U.S. The final adjustment to national income is the addition of depreciation, which is an income flow that is not included in U.S. national income. U.S. GDP using the income approach may thus be calculated as follows:

$$U.S.\ GDP = U.S.\ National\ Income - Net\ Foreign\ Factor\ Income + Depreciation$$

Figure 12.7 also allows us to see that net foreign factor income is the difference between Gross Domestic Product (GDP) and **Gross National Product (GNP)**. GNP was used more widely in the past in studies of aggregate output. It includes all the output produced and all the income earned by U.S. citizens whether working in the U.S. or abroad. In other words, GNP would equal A_1+A_2 in Figure 12.7, and so it would not include F_1.

GDP, on the other hand, would equal A_1+F_1 but it would not include A_2. The shift of focus from GNP to GDP in the U.S. occurred in the 1990s and makes sense in an increasingly globalized world where location seems more important than citizenship when thinking about contributions to the total production of the economy.

The income and expenditures approaches to the measurement of GDP should lead to the same numerical result because the approaches are based on the income-expenditure identity. The project of adding all the expenditure in the economy and all the income earned is such a massive project, however, that errors are inevitable. Because the two calculations do not match in practice, national income accountants include an item called **statistical discrepancy** to ensure that the two calculations are the same after accounting for the errors that arise from data collection. Table 12.5 shows the figures for U.S. GDP in 2017 using the income approach.

Table 12.5: The Different Components of Aggregate Income in the United States in 2017 (in billions of dollars)

	Amount (in billions of dollars)	Percentage of GDP
Employee Compensation	10,321.1	53.23%
Rent	743.9	3.84%
Interest	788.5	4.07%
Profits	3,268.7	16.86%
Taxes on Production and Imports (less subsidies)	1,268.8	6.54%
Depreciation	3,034.7	15.65%
Statistical Discrepancy	-35	-0.18%
Gross Domestic Product (GDP)	19,390.7	100%

Note: Some rounding error prevents the sum of the components of aggregate income from exactly equaling the total. It also prevents this GDP calculation from exactly equaling the GDP calculation in Table 12.4. Also, net foreign factor income is omitted because employee compensation, rent, interest, and profits already include the income of foreigners in the U.S. and exclude the income of Americans outside the U.S.

Source: U.S. Bureau of Economic Analysis. GDP and Personal Income: National Data. Table 1.10. Web. Last revised on March 28, 2018. Accessed on April 14, 2018. www.bea.gov.

A value for statistical discrepancy has been included so that the GDP calculation is the same (aside from rounding error) as the calculation using the expenditures approach in Table 12.4. To summarize the two approaches to GDP measurement in a single diagram, consider Figure 12.8.

Figure 12.8: A Diagram Representing the Two Approaches to GDP Measurement

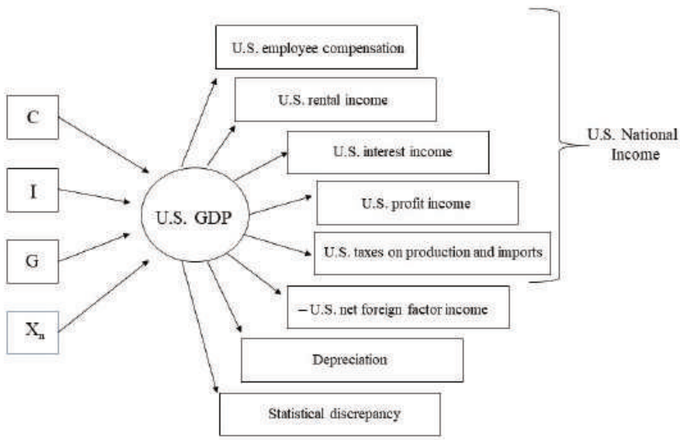


Figure 12.8 shows how the four major categories of aggregate expenditure generate incomes for workers, landlords, savers, and businesses (profits and depreciation funds) working and operating in the U.S. as well as U.S. federal, state, and local governments.

National income accountants use additional measures of macroeconomic performance. Starting with GDP, we can work backwards, in a sense, to obtain these other measures. For example, if we subtract depreciation from GDP, we obtain a measure called **Net Domestic Product (NDP)**, which is calculated as follows.

$$NDP = GDP - Depreciation$$

NDP shows us the value of the final goods and services produced in a year after we account for the depreciation of the capital stock. This measure addresses an issue raised earlier in this chapter regarding the possibility of double counting that results from the inclusion of final capital goods as an investment expenditure in the calculation of GDP. When a final capital good is used, it depreciates and adds value to the good that it is used to produce. If we count the full value of the final capital good in GDP and the value that it adds to final goods during the year, then double counting will occur. By subtracting depreciation, we eliminate the double counting problem. In 2017, U.S. NDP was calculated as follows:

$$U.S.\ NDP\ in\ 2017 = GDP - Depreciation = \$16,356\ billion$$

We can also reconstruct U.S. national income if we add net foreign factor income to U.S. NDP. This addition will add the incomes of Americans working abroad and subtract the incomes of foreigners working in the U.S.

$$NI = NDP + Net\ Foreign\ Factor\ Income$$

The GDP measure is a helpful way to think about aggregate output and aggregate income. Nevertheless, it is not a perfect measure. One of the problems with GDP is that it can change from year to year for reasons that do not seem to correspond with a change in the production of real final goods and services. For example, because prices are used as weights in the GDP measure, if all prices increase from one year to the next, then GDP will rise even if the real quantities produced have not changed. Let’s again consider a simple economy that only produces apples and oranges. Table 12.6 shows the quantities of each good and their prices for three different years.

Table 12.6: Nominal GDP in Three Different Years

Year	Apples	Price of apples	Oranges	Price of oranges	Nominal GDP
2001	100	$.50	200	$1.00	$250
2002	200	$1.00	400	$1.50	$800
2003	300	$1.50	600	$2.00	$1650

Nominal GDP is calculated by multiplying prices and quantities and summing them up as explained previously. **Nominal GDP** refers to GDP measured in current year market prices and is the measure that we have been discussing all along. As Table 12.6 shows, nominal GDP rose between 2001 and 2003, but it rose for two reasons. One reason is the increase in real quantities of apples and oranges produced. The second reason is the rise in the prices of apples and oranges. If we want our measure of aggregate output to only capture increases in real quantities produced, then we have a problem. The problem is the result of changing prices and so the obvious solution is to fix the prices. Which prices should we use? We have market prices from three different years that we can use in our calculation. It really does not matter which set of prices we use if they are constant. Because the choice is arbitrary, we will designate one year as the base year and then use the base year prices to calculate GDP for any year.

Table 12.7 shows how GDP is calculated using constant 2001 prices.

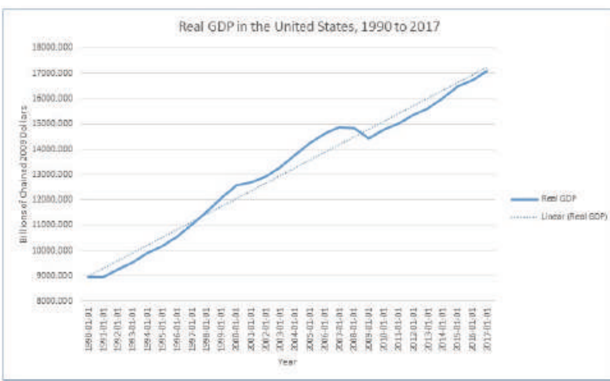
Table 12.7: Real GDP in Three Different Years

Year	Apples	Price of apples	Oranges	Price of oranges	Nominal GDP	Real GDP (2001 prices)
2001	100	$.50	200	$1.00	$250	$250
2002	200	$1.00	400	$1.50	$800	$500
2003	300	$1.50	600	$2.00	$1650	$750

That is, 2001 has been selected as the base year and has been used to calculate real GDP. **Real Gross Domestic Product (GDP)** is the measure of aggregate output using current year quantities and base year or constant prices. It should be clear that Real GDP has risen between 2001 and 2003 but the increase is much less dramatic than the increase in nominal GDP during those years. The reason, of course, is that real GDP only increases because of the increase in real quantities. Nominal GDP, on the other hand, rises due to the increase in real quantities and the rise in market prices. Because Real GDP allows us to focus on changes in real production, it is argued to be a superior measure of aggregate output.

During the past century, real GDP has increased dramatically in the United States. When economists talk about **economic growth**, they have in mind a rise in real GDP. The growth of real GDP has followed an upward long-term trend but the growth has not been smooth. At various times, real GDP has fallen and at other times, it has risen dramatically. These booms and busts of the economy are referred to as the **business cycle**. Figure 12.9 shows how real GDP has fluctuated around a long-term trend line representing economic growth between 1990 and 2017.[18]

Figure 12.9: Real GDP Fluctuations Around a Long-Term Trend Line, 1990 to 2017 (in billions of 2009 chained dollars)

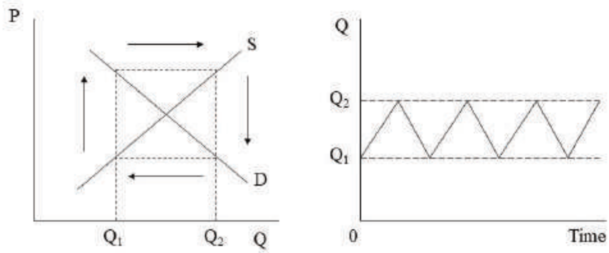


Note: Real GDP is measured on an annual basis and is not seasonally adjusted.

Reductions in real GDP are considered **recessions** and increases in real GDP are considered **expansions**. When real GDP reaches a local maximum, it is considered the **peak** of the business cycle. When real GDP reaches a local minimum, it is considered the **trough** of the business cycle. As a rule of thumb, economists usually regard two consecutive quarters of negative real GDP growth as a recession. Recessions vary a great deal, however, in terms of their length and intensity. They can be long and deep or short and mild. The length and intensity of expansions is also difficult to predict.

Later chapters are devoted to theoretical explanations of business cycle fluctuations, but we can already begin to see how macroeconomic cycles may have their roots in fluctuations at the microeconomic level. Consider the supply and demand model as it pertains to competition in the market for a single good or service. When the price is above the equilibrium price, a surplus exists. Competition then drives the price downward. If the price falls so much that it overshoots the equilibrium price, then a shortage will emerge. Competition will then drive the price upwards, which might overshoot the equilibrium price again. As the price falls and then rises again, the quantity supplied falls and then rises again. The **overshooting** of the equilibrium price thus gives rise to a series of decreases and increases in the quantity available for purchase in the marketplace. Figure 12.10 represents the series of surpluses and shortages that emerge in the market and how the output of this product that is available in the market fluctuates over time.

Figure 12.10: Structural Cycles: The Cobweb Model



This simple model is referred to as the **cobweb model**.[19] More sophisticated versions of the cobweb model show the market price gradually approaching the equilibrium price as the adjustments become smaller and begin to approach their target. Assuming many or all individual markets experience such fluctuations, an aggregation of these output fluctuations will produce macroeconomic fluctuations with corresponding fluctuations in employment. Because the explanation of business fluctuations stemming from the cobweb model is rooted in errors made by producers, it may be considered a heterodox theory of economic cycles. As we will see, however, most orthodox and heterodox explanations of business cycles emphasize other factors. Later chapters delve into the sources of these different explanations.

In addition to real GDP, economists also frequently discuss the growth rate of real GDP. The **real GDP growth rate** is calculated as follows:

$$Real\ GDP\ growth\ rate = \frac{Real\ GDP_t - Real\ GDP_{t-1}}{Real\ GDP_{t-1}}$$

The calculation of the real GDP growth rate between year t-1 and year t divides the change in real GDP by the real GDP of the previous year. A positive rate of real GDP growth suggests that real GDP has increased since the previous year. A negative rate of real GDP growth suggests that real GDP has fallen since the previous year. A zero rate of real GDP growth suggests that real GDP has remained the same since the previous year.

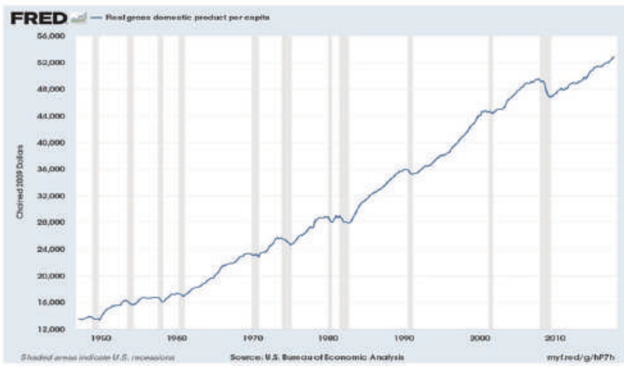
Another measure that economists use is **per capita real GDP** or real GDP per person. Per capita real GDP is calculated as follows:

$$Per\ capita\ real\ GDP = \frac{Real\ GDP}{Population}$$

Per capita real GDP is often considered to provide a rough measure of the standard of living in a nation. It measures real income or real output per person. The measure has a serious shortcoming, however, because it is only an average and does not tell us anything about an individual's economic welfare. If everyone receives the same real income, then per capita real GDP tells us what that real income level is. If a great deal of income inequality exists, however, then some people will have real incomes that are far below the per capita real GDP. Other people will have real incomes that are far above the per capita real GDP. In other words, per capita real GDP tells us nothing about the distribution of income. If someone suggests that it is a rough measure of what individuals earn in real terms, then that suggestion can be very misleading.

Nevertheless, a rise in per capita real GDP gives us a sense of how much the economy has expanded over time. Figure 12.11 shows how the per capita real output has increased dramatically from 1947 to 2017.

Figure 12.11: Real Gross Domestic Product Per Capita, 1947-2017



Source: U.S. Bureau of Economic Analysis, Real gross domestic product per capita [A939RX0Q048SBEA], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/A939RX0Q048SBEA, April 14, 2018.

A similar measure that uses real GDP in its calculation is **real GDP per worker**, which measures the average labor productivity for each member of the labor force as follows.

$$Real\ GDP\ per\ worker = \frac{Real\ GDP}{Labor\ Force}$$

A higher level of real output per worker suggests higher average labor productivity. A lower level of real output per worker suggests lower average labor productivity. Again, this measure is subject to the same shortcoming in that it is only an average. Nevertheless, economists widely use both the labor productivity and per capita output measures. Economists also refer to the growth rates of real per capita income and real output per worker as measures of economic growth and productivity changes. A productivity growth slowdown began in the 1970s. The reasons for the productivity slowdown are hotly debated. Some explanations focus on the inflation that occurred in the 1970s while other explanations focus on the breakdown of the cooperative labor-management relations of the postwar period. The macroeconomic theories in the second part of this book offer explanations for such changes.

Heterodox Critiques of National Income Accounting

Many heterodox economists are sharply critical of national income accounting. This section concentrates on two major critiques of GDP as a measure of economic well-being. The first critique is one that feminist economists have developed to draw attention to the many contributions that women make to our economic well-being that have been excluded from the calculation of GDP. The second critique involves the assertion that human happiness depends on more than the amount of goods and services that are available for consumption. We consider each critique separately.

GDP only includes the market value of all final goods and services produced within the economy during a given year. Because it only includes market values, any production that never finds its way to the market is necessarily omitted from the calculation of GDP due to the lack of a market price. One type of production that never enters the market is household production. Historically, women have been the primary producers within the home of a huge variety of goods and services, including home-cooked meals, laundry services, cleaning services, childcare, care of elderly family members, care of pets, transportation for children and the elderly, gardening and landscaping, clothing, clothing repair, and grocery shopping. The list could easily be expanded.

When people outside the home are hired to perform these services and produce these goods, the production is counted in GDP because it has a market value. When women perform these duties within the home and their families consume the goods and services, they are not counted in GDP. Consequently, an enormous amount of labor that women have performed during the past century has been completely overlooked in the national income accounts. It is the invisible nature of women's unpaid work in the home that has been the motivation for sharp feminist critiques of GDP as a measure of economic welfare.

Some early estimates of national income in Norway and other Scandinavian countries included the value of unpaid household labor. For example, in Norway in 1943, the value of unpaid household labor was estimated to be 15% of national product.[20] As the United Nations prepared to introduce the first international standard for national accounts in 1953, however, goods and services derived from unpaid household work were excluded, which led Norway to eliminate it from its national accounts in 1950.[21]

The primary method of accounting for goods and services that do not have market values is to use imputed values. The **imputed value** of a good or service is based on its likely value in the market if it was sold. It can be thought of as the opportunity cost of consuming the good or service when it could be sold. The practical way to handle this problem when considering unpaid labor in the home is to assume that "an hour of market work and an hour of nonmarket work have the same value."[22] The market wage of a substitute household worker is then used in combination with the labor time needed to produce goods and services in the home.[23] The typical choice of yardstick to arrive at the estimates of the value of household labor is **extra gross wages**, which are before taxes and include employers' social security contributions.[24] A rough estimate of the value of nonmarket household production is approximately half of GDP in the industrialized countries.[25] Given how massive this contribution is, the omission of unpaid household labor in the national income accounts grossly understates our national output of goods and services.

At the same time, feminist economists recognize that many activities within the home have an intrinsic value that market values simply cannot capture.[26] To represent these intrinsic values, quality of life measures are used that include the "pursuit of good health, the acquisition of knowledge, the time devoted to fostering social relationships, [and] the hours spent in the company of relatives and friends."[27] However we might measure the contribution of unpaid household labor, it is essential to recognize that it is *women's* contribution that is mainly being overlooked in the national income accounts. In industrialized nations in recent decades, women spent about 2/3 of their total work time on unpaid nonmarket activities and 1/3 of their time on paid market activities whereas for men the shares have been reversed.[28] In developing nations, the difference between men's and women's shares is even greater.[29]

For an alternative measure of macroeconomic well-being, we turn to the Himalayan Mountains where Bhutan's primary measure is something called **Gross**

National Happiness (GNH). GNH has become a guiding light of economic policymaking in Bhutan. Since Bhutan became a democracy in 2008, its Constitution has required its leaders "to consult the four pillars of Gross National Happiness – good governance, sustainable socioeconomic development, preservation and promotion of culture, and environmental conservation – when considering legislation."[30]

Bhutan's rejection of GDP as a measure of economic progress goes back to 1971.[31] This Buddhist approach to economic well-being places emphasis on "the spiritual, physical, social and environmental health of its citizens and natural environment."[32] To protect the natural environment, Bhutan has taken extraordinary measures. It is committed to remaining carbon neutral and to permanently maintaining 60% of its landmass under forest cover, which has included a ban on export logging.[33] The GNH concept has also influenced Bhutan's system of education, which places heavy emphasis on environmental protection, recycling, and daily meditation.[34]

King Jigme Singye Wangchuck, who ruled Bhutan until 2006, coined the GNH label decades ago.[35] It is worth noting that the concept was developed in Asia rather than in western nations where material possessions have long served as the measure of well-being. Nevertheless, the concept has caught on with western leaders. In 2011, the UN General Assembly "passed a resolution inviting member states to consider measures that could better capture 'the pursuit of happiness' in development," which led to the first World Happiness Report of 2012.[36] The UN uses a variety of different variables to calculate a score for each country, which serves as its index of happiness. According to the 2018 World Happiness Report, Finland is the happiest nation on the planet and the U.S. has fallen to 18th place due to crises of obesity, substance abuse, and mental health problems.[37] Although GNH has not replaced GDP as the measure of greatest interest to most economists, it has drawn public attention to the possibility that our economic welfare depends on more than the amount of final goods and services our nation produces each year.

The Measurement of the Labor Force and the Unemployment Rate

We now turn to the measurement of the labor force, employment, and unemployment. Within the U.S. Department of Labor, the Bureau of Labor Statistics (BLS) is responsible for publishing the unemployment rate each month. To understand this calculation and related measures, consider Figure 12.12, which breaks down the population into its component parts.

Figure 12.12: The Breakdown of the Population Using Labor Force Aggregates

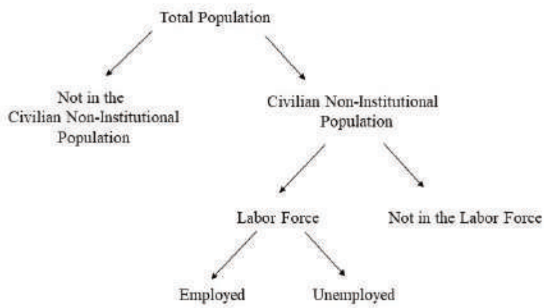


Figure 12.12 shows that the total population may be divided into the civilian non-institutional population and those not in the civilian non-institutional population. Those in the **civilian non-institutional population** are at least 16 years old and are not in institutions such as mental hospitals or prisons. Those **not in the civilian non-institutional population** are under 16 years of age or are living in institutions. The civilian non-institutional population then may be divided into those in the labor force and those not in the labor force. Those in the **labor force** are non-institutionalized civilian workers who are willing and able to work. Those **not in the labor force** are non-institutionalized civilian workers who are not willing or are not able to work. For example, full-time students, retirees, stay-at-home parents, and disabled people are considered not in the labor force. They are of working age but are not willing or able to work at the current time. Of those willing and able to work, those with jobs are considered **employed**. Those without jobs

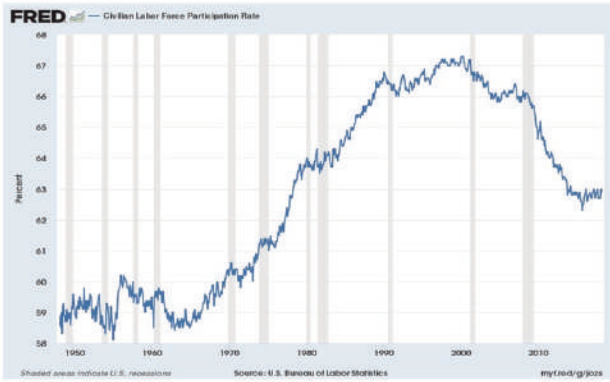
who have tried to find work within the past four weeks are considered **unemployed**. If workers have become discouraged and have given up looking for work, then they are considered outside the labor force. **Discouraged workers** are thus not in the labor force. The four-week cutoff is completely arbitrary and shows how social values creep into the construction of macroeconomic variables like unemployment. That is, some people find this cutoff to be too short when counting people as unemployed. These people wish to count more people as unemployed. Other people find this cutoff to be too long when counting people as unemployed. These people wish to count fewer people as unemployed. The normative content of the unemployment measure was discussed in detail in Chapter 1.

Given these definitions, we can now construct two key labor force measures: 1) the labor force participation rate and 2) the unemployment rate. The **labor force participation rate** is the labor force divided by the civilian non-institutional population. It shows us the fraction of the civilian non-institutional population that is willing and able to work and is calculated as follows:

$$Labor\ Force\ Participation\ Rate = \frac{Labor\ Force}{Civilian\ Non\text{-}institutional\ Population}$$

Figure 12.13 shows the pattern of the labor force participation rate in the United States from 1948 to 2018.

Figure 12.13: The U.S. Civilian Labor Force Participation Rate, 1948-2018



Source: U.S. Bureau of Labor Statistics, Civilian Labor Force Participation Rate [CIVPART], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/CIVPART, April 14, 2018.

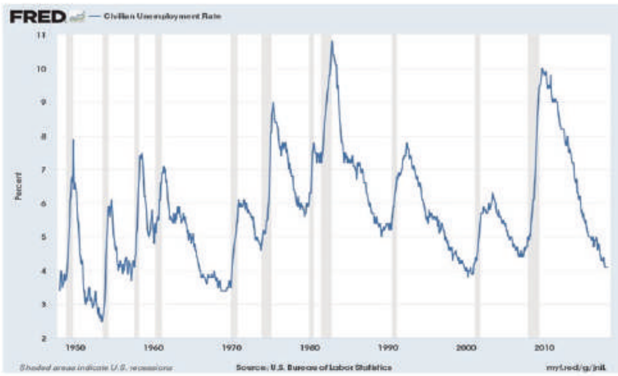
Figure 12.13 shows that the labor force participation rate in the U.S. rose considerably from the 1960s to the 1990s but that it has declined quite significantly since the turn of the century.

The **unemployment rate** is the percentage of the labor force that is unemployed. It is calculated as follows:

$$Unemployment\ Rate = \frac{Unemployed}{Labor\ Force}$$

Figure 12.14 shows that the unemployment rate in the U.S. has fluctuated considerably during the past century but that it has not followed an obvious upward or downward trend. It should also be clear that some unemployment has always existed, and it never seems to approach zero.

Figure 12.14: The U.S. Civilian Unemployment Rate, 1948-2018



Source: U.S. Bureau of Labor Statistics, Civilian Unemployment Rate [UNRATE], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/UNRATE, April 14, 2018.

The U.S. labor force figures for December 2017 are below, and they have been used to calculate the unemployment rate and the labor force participation rate.[38]

$$Total\ Population = 325,719,178$$

$$Civilian\ Non-institutional\ Population = 256,109,000$$

$$Not\ in\ the\ Civilian\ Non-institutional\ Population = 69,610,178$$

$$Labor\ Force = 160,597,000$$

$$Not\ in\ Labor\ Force = 95,512,000$$

$$Employed = 154,021,000$$

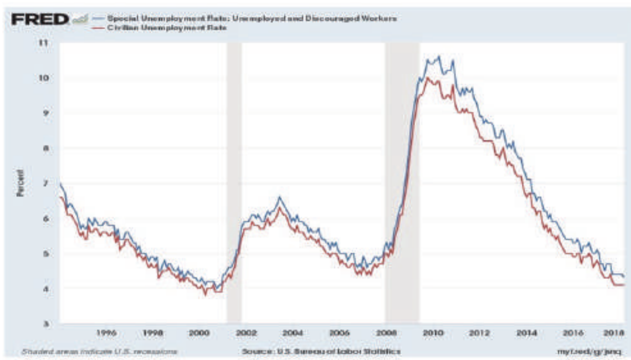
$$Unemployed = 6,576,000$$

$$Labor\ Force\ Participation\ Rate = 62.7\%$$

$$Unemployment\ Rate = 4.1\%$$

Although the unemployment rate gives us some idea as to the percentage of workers who want jobs but are unable to find them, it has some shortcomings. As we have seen, it excludes discouraged workers and so it tends to understate the amount of unemployment. Figure 12.15 provides an historical look at the pattern of the **special unemployment rate** since 1994 in the U.S.

Figure 12.15: The U.S. Special Unemployment Rate: Unemployed and Discouraged Workers, 1994-2018



Source: U.S. Bureau of Labor Statistics, Special Unemployment Rate: Unemployed and Discouraged Workers [U4RATE], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/U4RATE, April 14, 2018.
U.S. Bureau of Labor Statistics, Civilian Unemployment Rate [UNRATE], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/UNRATE, April 14, 2018.

The special unemployment rate includes both the officially unemployed and the discouraged workers. Figure 12.15 shows that the inclusion of discouraged workers in the measurement of the special unemployment rate causes the special unemployment rate to be significantly higher than the official unemployment rate at any given time.

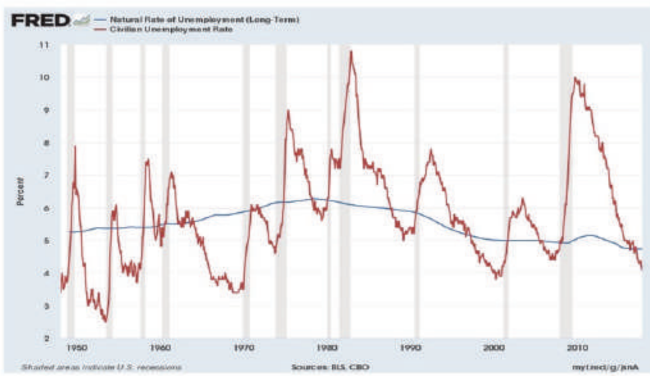
The unemployment rate is also based on a headcount and treats part-time workers as employed even if they would like to have full-time work, which understates the amount of unemployment in the economy. Finally, it might overstate the amount of unemployment if people indicate in the survey that they have tried within the past four weeks to find work because they believe that it will help them qualify for unemployment benefits.

The causes of unemployment are the focus of later chapters because identifying them requires the use of theoretical frameworks. At this point, however, it is worth summarizing the neoclassical perspective on unemployment, which helps us grasp a major source of the theoretical disagreements that we will encounter in later chapters. Neoclassical economists argue that some unemployment is inevitable and perfectly acceptable in a market capitalist economy. This type of unemployment is called **natural unemployment**. Furthermore, it consists of two types of unemployment: structural unemployment and frictional unemployment. **Structural unemployment** occurs when workers lose their jobs due to shifts of consumer demand or technological changes. Such shifts are inevitable and necessary within capitalist economies, so the argument goes, because consumers are free to make choices and firms are free to introduce new technologies. When such changes occur, some industries decline as other industries expand. The result is that workers in contracting industries will become unemployed as they strive to find jobs in the expanding industries. **Frictional unemployment** occurs when new entrants into the labor force are looking for that first job or when they voluntarily decide to leave one employer and find another employer for whom to work. These increases in unemployment are considered unavoidable in an unregulated economic system where workers are free to make their own decisions about employment. Using these two measures of unemployment, we can define a natural rate of unemployment (NRU).

$$NRU = \frac{Structurally\ Unemployed + Frictionally\ Unemployed}{Labor\ Force}$$

The NRU is the subject of debate among neoclassical economists. How much of the unemployment that we observe results from these factors? The U.S. Congressional Budget Office has estimated the NRU for different years. If we place the NRU on a graph of the official unemployment rate for the period 1948-2017, then we obtain Figure 12.16.

Figure 12.16: The U.S. Natural Rate of Unemployment and the Civilian Unemployment Rate, 1948-2017



Source: U.S. Congressional Budget Office, Natural Rate of Unemployment (Long-Term) [NROU], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/NROU, April 14, 2018.
U.S. Bureau of Labor Statistics, Civilian Unemployment Rate [UNRATE], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/UNRATE, April 14, 2018.

Figure 12.16 shows that a significant amount of unemployment has existed beyond the NRU in recent decades. This amount of unemployment beyond the NRU is labeled **cyclical unemployment**. As Figure 12.16 shows, cyclical unemployment was especially high during the recessions of the early 1980s and during the Great Recession. It is argued to stem from avoidable factors. It rises and falls according to the phase of the business cycle. For neoclassical economists, it is the only source of concern when considering the unemployment problem. At times, the unemployment rate has fallen so low that it falls below the NRU. When unemployment falls to such low levels, it means that hiring is happening at such a furious pace that even the structurally unemployed and the frictionally unemployed are being hired. If the unemployment remains at this low level for a long enough period, then economists might revise their estimate of the NRU in a downward direction. Such revisions to the NRU have occurred in the past as Figure 12.16 indicates.

Because of their division of unemployment into natural unemployment and cyclical unemployment, neoclassical economists mean something very peculiar when they refer to full employment. **Full employment** for neoclassical economists does not mean a zero rate of unemployment. When the economy reaches full employment, it means that cyclical unemployment is zero and the unemployment rate is equal to the NRU. In other words, unemployment exists but it is only structural unemployment and frictional unemployment. If the unemployment rate falls below the NRU as previously discussed, then the economy operates at a level beyond full employment. Finally, the level of real GDP corresponding to full employment is the **potential GDP** of the economy. If the economy is at the potential GDP, then it is operating on its production possibilities frontier (PPF). All resources are fully employed using the most efficient methods of production.

The Measurement of the Aggregate Price Level and the Inflation Rate

So many different goods and services exist within an economy that it is difficult to think about something called the general level of prices. Nevertheless, orthodox and heterodox economists devote a lot of attention to it. To measure the general price level, it is necessary to use what economists call a price index. A **price index** is a summary measure or statistic that is supposed to measure the general price level. When it changes from one period to the next, the change is supposed to capture changes in many different prices at once. Economists use many different price indices to measure the general price level. In this chapter, we will concentrate on two price indices that are the most widely used measures of the general price level.

The first measure of the general price level is the GDP deflator. The **GDP deflator** is calculated as the ratio of nominal GDP to real GDP as follows:

$$GDP\ Deflator = \frac{Nominal\ GDP}{Real\ GDP}$$

Table 12.8 shows how the GDP deflator is calculated for four years using 2001 as the base year.

Table 12.8: The GDP Deflator (an index of prices)

Year	Nominal GDP	Real GDP (2001 prices)	GDP Deflator
2001	$250	$250	1.00
2002	$800	$500	1.60
2003	$1650	$750	2.20
2004	$1900	$1000	1.90

This ratio might seem like a strange measure of the general price level, but consider what causes a divergence between nominal GDP and real GDP. The only reason that nominal GDP exceeds real GDP in a specific year is that the price level has risen. If the GDP deflator is equal to 1.60 in 2002 and 2001 is the base year, then the implication is that the general price level is 160% of what it was in the base year. A GDP Deflator of 2.20 in 2003 implies that the general price level is 220% of what it was in the base year. In general, a rise in the GDP deflator means that the general price level has risen.

Using the GDP deflator, it is possible to calculate the rate of inflation. The **inflation rate** measures the percentage change in the general price level from one year to the next. To calculate the inflation rate, a price index (P) is required. Using any price index, we can calculate the inflation rate in year t as follows:

$$Inflation\ rate = \frac{P_t - P_{t-1}}{P_{t-1}}$$

In words, if we divide the change in the general price level since the previous year by the previous year's price level, then we obtain the percentage change in the general price level since the previous year. The inflation rate thus tells us how rapidly prices are rising. Since the GDP deflator is a price index, we can use it to calculate the inflation rate. For example, the inflation rate in 2003 would be calculated as follows using the information from Table 12.8.

$$Inflation\ rate = \frac{P_{2003} - P_{2002}}{P_{2002}} = \frac{2.2 - 1.6}{1.6} = 37.5\%$$

Table 12.9 adds the inflation rates to the information in Table 12.8.

Table 12.9: The Rate of Inflation

Year	Nominal GDP	Real GDP (2001 prices)	GDP Deflator	Inflation Rate
2001	$250	$250	1.00	--
2002	$800	$500	1.60	60%
2003	$1650	$750	2.20	37.5%
2004	$1900	$1000	1.90	-13.64%

In Table 12.9 the inflation rate for 2001 is shown as undefined. It is only because the inflation rate for 2000 is not included in the table that we cannot calculate the inflation rate for 2001. The reader should notice that the inflation rate for 2004 is negative because the price level fell relative to 2003. **Deflation** is the name that economists use to describe a negative rate of inflation.

The reason that the GDP deflator is referred to as a deflator is that it makes it possible to deflate nominal magnitudes to obtain real magnitudes. For example, suppose that we know the nominal GDP in 2002 is $800 and the GDP deflator is 1.6. Using the definition of the GDP deflator, we can deflate the nominal GDP and solve for the real GDP in 2002 as follows:

$$Real\ GDP\ in\ 2002\ in\ 2001\ dollars = \frac{Nominal\ GDP}{GDP\ Deflator} = \frac{\$800}{1.60} = \$500$$

Using the deflator, we eliminate the impact of the rising price level to express GDP in real terms (i.e., measured in constant, base year dollars).

A second measure of the general price level upon which economists rely heavily is the consumer price index (CPI). The Bureau of Labor Statistics (BLS) within the U.S. Department of Labor publishes the CPI each month. Unlike the GDP deflator which includes the prices of all final goods and services produced in the economy, the CPI only includes the prices of goods and services that a typical consumer purchases. In fact, it is based on the price of a typical consumer basket of goods and services. The information used to construct the CPI is derived from the Consumer Expenditure Survey, which is administered to thousands of families in the United States each year. This information helps the BLS determine which consumer goods and services American households purchase. The BLS then collects information on thousands of prices of goods and services each month to construct the CPI. The BLS uses eight major categories of expenditure to organize the items contained in the consumer basket that it uses, which include the following:[39]

1. Food and beverages (breakfast cereal, milk, coffee, chicken, wine, service meals and snacks)
2. Housing (rent of primary residence, owners' equivalent rent, fuel oil, bedroom furniture)
3. Apparel (men's shirts and sweaters, women's dresses, jewelry)
4. Transportation (new vehicles, airline fares, gasoline, motor vehicle insurance)
5. Medical Care (prescription drugs and medical supplies, physicians' services, eyeglasses and eye care, hospital services)

6. Recreation (televisions, pets and pet products, sports equipment, admissions)
7. Education and communication (college tuition, postage, telephone services, computer software and accessories)
8. Other goods and services (tobacco and smoking products, haircuts and other personal services, funeral expenses)

Within each category, the BLS tracks the prices of hundreds of representative items and uses them to calculate the CPI. To see exactly how the CPI is calculated, let's consider a simple example in which only two goods are included in the typical consumer's basket of goods. Table 12.10 shows hypothetical price information for five years.

Table 12.10: Calculating the CPI: A Hypothetical Example (base year = 2007)

Year	Units of Food	Price of Food	Units of Fuel	Price of Fuel	Price in the base year	Price in the current year	CPI	Inflation Rate (%)
2007	400	15	750	8	12,000	12,000	1.00	--
2008	400	18	750	9	12,000	13,950	1.163	16.25
2009	400	19	750	12	12,000	16,600	1.383	19.00
2010	400	23	750	14	12,000	19,700	1.642	18.67
2011	400	20	750	13	12,000	17,750	1.479	-9.90

In Table 12.10, the typical consumer purchases 400 units of food and 750 units of fuel. Food and fuel are thus the only two goods in the consumer's basket of goods. The base year is designated as 2007 and so the BLS has decided that these quantities of food and fuel were the quantities that a typical consumer purchased in 2007. It is important to notice that these quantities are then held constant for every other year. It is possible that a consumer might alter the mix of goods and services in her basket, but the BLS assumes a fixed basket for every year when constructing the CPI. The prices of food and fuel, on the other hand, change from year to year (or month to month in reality), and the BLS tracks these changes closely.

The next step is to calculate the price of the consumer basket of goods and services in the base year. This calculation is carried out by simply multiplying the price in 2007 (the base year) times quantity consumed in 2007 for each good and adding them up to obtain $12,000. Since the price of the basket in the base year is used to calculate the CPI in each year, the 12,000 has been included as a fixed value in a single column. We then calculate the price of the consumer basket in the current year by multiplying the fixed quantities and the current year prices of the goods and adding them. Because the prices in the current year change from year to year, the price of the basket in the current year changes too.

The final step in the calculation of the CPI for each year is to divide the price of the basket in the current year by the price of the basket in the base year as follows:

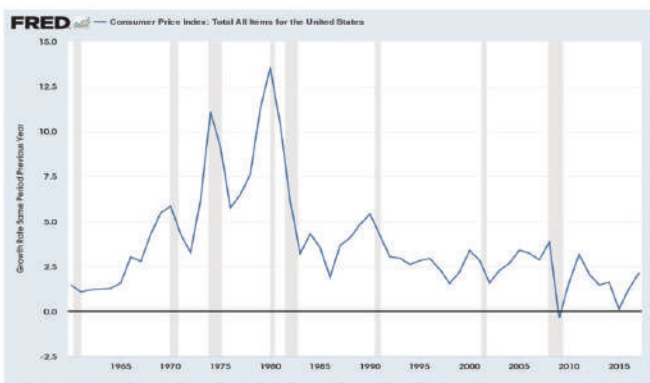
$$CPI = \frac{Price\ of\ basket\ in\ the\ current\ year}{Price\ of\ basket\ in\ the\ base\ year}$$

Looking at how the CPI has changed from year to year shows us how the general price level rose from 2007 to

2010 and then fell in 2011. It is also possible to calculate the annual inflation rate, using the CPI values in the formula for the inflation rate provided earlier. The inflation rates have also been included in Table 12.8. The decline in the price level between 2010 and 2011 has caused the inflation rate to turn negative, which indicates deflation, as previously explained.

Figure 12.17 shows how the CPI inflation rate has fluctuated during the past half century.

Figure 12.17: The U.S. CPI Inflation Rate, 1960-2017



Source: Consumer Price Index: Total All Items for the United States [CPALTT01USA659N], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/CPALTT01USA659N, April 14, 2018.

Noteworthy periods include the double-digit inflation rates of the 1970s when the oil price shocks occurred and the low inflation rates of the 1990s when the economy expanded and the introduction of new technologies led to rising productivity.

The CPI is a helpful index of the general price level. The CPI inflation rate is used to adjust the poverty thresholds discussed earlier in this chapter. Nevertheless, it has been recognized for its limitations, even among mainstream economists.[40] The main limitation stems from the likely possibility that consumers will change the mix of goods and services they buy over time. For example, when the price of a good in the basket rises, consumers may substitute a lower-priced good outside the basket for the higher priced good. Because the consumer basket remains fixed, however, the CPI will register a larger price increase than can be justified by the consumer's purchasing behavior. Only a change in the consumer basket can rectify this so-called **substitution bias**, which causes increases in the CPI to overstate the increase in the cost of living.

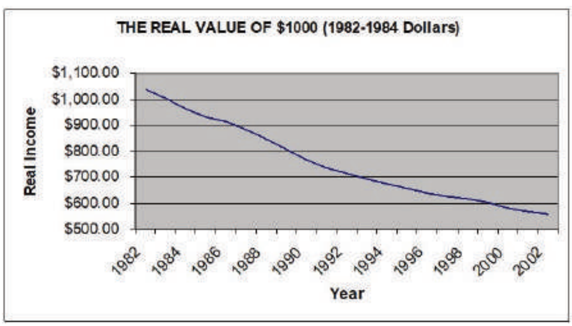
Another kind of bias is the so-called **new goods bias**. When new goods become available in the marketplace, consumers might purchase them even though they are not in the fixed basket of goods. Changes in the CPI will not reflect changes in the prices of these goods, which leads to an inaccurate measure of the cost of living. Finally, a **quality change bias** makes the CPI less reliable than it would otherwise be. When the quality of a good in the basket increases along with its price, the CPI will only register the price increase. It will not correct for the fact that the consumer acquires a higher quality product for her money. The result is a tendency for the CPI to overstate the rise in the price level.

Just like the GDP deflator, it is possible to use the CPI to deflate nominal dollar amounts into real dollar amounts. To make such conversions, we simply divide the nominal dollar amounts by the CPI as follows:

$$Real\ income = \frac{Nominal\ income}{CPI}$$

For example, suppose that you have $1000 in the early 1980s that you place under your mattress at home.[41] You then pull out the money in 2002 and realize that your money is worth a lot less than it was back in the early 1980s. Why? The reason is that the prices of goods and services have increased considerably during that period, which has eroded the real value of the $1000. Figure 12.18 shows how the real value of the $1000 fell as the price level rose.

Figure 12.18: The Effect of Inflation on the Real Value of $1000



The base year in this example is 1982-1984. Actual CPI data were used for these calculations and so this decline in the real value of the $1000 is not hypothetical but actual. As Figure 12.18 shows, the real value of the $1000 in 2002 was equivalent to about $550 in 1982-1984 dollars. Therefore, it lost nearly half its value (i.e., its purchasing power). For this reason, people generally try to protect themselves from inflation. People who receive fixed incomes or people who put money under mattresses generally are harmed by inflation. It explains why workers organize themselves into unions to pressure employers for higher money wages and why Social Security recipients are protected with automatic cost of living adjustments to their benefit amounts.

Inflation may also harm moneylenders if they do not charge sufficient interest to compensate them for the rising price level. Consider how we might approximate the percentage change in the real purchasing power of a sum of money (M):

$$\%\Delta \frac{M}{P} \approx \%\Delta M - \%\Delta P$$

If a moneylender lends M to a borrower, then the percentage change in M/P is the real interest rate, where P represents the price level. The **real interest rate** is the percentage change in the purchasing power of a loan and represents the real interest that the lender receives from the borrower. It may be approximated as the difference between the percentage change in the nominal amount of the loan and the percentage change in the price level. Since the percentage change in the nominal amount of the loan is the **nominal interest rate (i)** and the percentage change in the price level is the inflation rate (π), we can approximate the real interest rate (r) in the following way:[42]

$$r \approx i - \pi$$

Therefore, the real interest rate that a lender will earn is equal to the nominal interest rate minus the inflation rate. For example, if a lender makes a loan and charges 10% interest but the inflation rate is 4%, then the real interest rate is only 6%. The nominal dollar amount of the loan grew by 10%, but the 4% rise in the price level

caused the lender's purchasing power to only rise by 6%. The equation can also be rearranged as follows:
$i \approx r + \pi$

This equation suggests that a moneylender can protect himself if he charges a nominal interest rate that is equal to the inflation rate plus whatever real return he wants to earn. In this way, interest rates can be set to compensate the lender for whatever inflation is expected to occur. Of course, the nominal interest rate and thus the real interest ultimately will be decided in the market for loanable funds, but inflation will influence how much lenders wish to lend at each nominal interest rate.

The approximation of the real interest rate shows us that an unexpected surge in inflation can harm a lender, especially if the inflation rate rises above the nominal interest rate. In that case, the real interest rate will be negative and the lender will lose purchasing power between the time the loan is made and the time that it is repaid. Borrowers benefit in this case because they pay less in real terms for the loan. If inflation is unexpectedly low, however, then the lender will enjoy a benefit because the real interest rate will be higher than expected. Borrowers are harmed in this case because they pay more in real terms for the loan.

Despite these issues, many economists do not worry too much about low or moderate amounts of inflation because inflation affects both nominal incomes and the price level. Therefore, real incomes are often left unaffected when inflation is relatively low. Extremely high rates of inflation called **hyperinflation** can cause economic turmoil, however, and are regarded as very damaging to a nation's economic well-being. In those situations, money loses its value so rapidly that it ceases to effectively serve as a medium of exchange. Desperate for a stable measure of value in individual exchanges, people faced with hyperinflation often abandon the currency and resort to barter exchange.

Following the Economic News[43]

The GDP estimates are frequently revised for many months and even years later. Such adjustments are inevitable given the complexity of the project. According to *The Wall Street Journal*, the Commerce Department revised its GDP growth estimate for the fourth quarter of 2017 upwards to 2.9% in March 2018. The previous estimate had been 2.5%, which Ben Leubsdorf explains was weaker growth than occurred in the second and third quarters of 2017. The WSJ writer explains that the rise in GDP was the result of a 4% annual rate increase in personal consumption expenditures, which in turn was driven by a strong increase in the demand for durable goods. Investment expenditures also rose, reports Leubsdorf, including a 6.8% increase in business fixed investment, a 11.6% rise in expenditures on final capital goods, and a 12.8% increase in residential investment expenditures. Government spending was also up 3.0% at an annual rate in the fourth quarter of 2017. The WSJ attributes the weaker GDP growth to the trade deficit and to a negative change in inventories. All these factors relate to the expenditures approach, but Leubsdorf also refers to factors that are relevant to the income approach to the measurement of GDP. The weaker GDP growth relative to the previous quarters is attributed in part to the reduction in the corporate income tax rate from 35% to 21% as of January 1, 2018. Some corporations found an incentive in this change in the tax code to "shift earnings and expenses" from late 2017 to early 2018. As a result, the reported profits were lower in the fourth quarter of 2017 than they would have been. The consequence of a reduction in reported profit income in the fourth quarter of 2017, of course, is a lower GDP in that quarter. This reduction was apparently offset by other components of aggregate income, but the weak profits caused economic growth to be slower overall. The benefit of understanding national income accounting is that one can read about the various factors that affect GDP and organize those factors into expenditure side and income side components.

Summary of Key Points

1. The official poverty rate measures the percentage of the population that lives below the poverty threshold.
2. The quintile ratio measures the ratio of the income of the top 20% of the population to the income of the bottom 20% of the population.
3. The Gini coefficient is a measure of income inequality (or wealth inequality) that ranges from 0 to 1 with 0 representing perfect income (wealth) equality and 1 representing perfect income (wealth) inequality.
4. Gross domestic product (GDP) measures the market value of all final goods and services produced within the geographic boundaries of a nation within a given year and excludes the values of intermediate goods, transfer payments, purchases of financial assets, and the values of used goods.
5. The expenditures approach to the measurement of GDP involves the sum of consumer spending,

investment spending, government spending, and net export spending.

6. The income approach to the measurement of GDP involves the sum of U.S. employee compensation, U.S. rental income, U.S. interest income, U.S. profit income, taxes on production and imports, net foreign factor income, depreciation, and statistical discrepancy.
7. Nominal GDP measures aggregate output using current market prices whereas real GDP measures aggregate output using constant, base year market prices.
8. Feminist economists are critical of the GDP measure because it excludes unpaid household work, which women have mainly performed. The use of imputed values to estimate women's unpaid contribution only addresses part of the problem due to the intrinsic value of much of this work that market values do not reflect.
9. Gross National Happiness (GNH) is an alternative measure of economic well-being that is based on spiritual, physical, social, and environmental health.
10. The official unemployment rate measures the percentage of the labor force that is unemployed. The labor force participation rate measures the percentage of the civilian non-institutional population that is in the labor force.
11. The GDP deflator is the ratio of nominal GDP to real GDP and serves as a measure of the general price level. It may also be used to calculate the inflation rate.
12. The Consumer Price Index (CPI) measures the value of a basket of consumer goods and may be

used to estimate the general price level and to calculate the inflation rate.

13. The real interest rate measures the real cost of borrowing and is calculated as the difference between the nominal interest rate and the inflation rate.

List of Key Terms

Poverty threshold

Ratio of income to poverty

Poverty

Near poverty

Deep poverty

Income deficit

Income surplus

Official poverty rate

Countercyclical

Supplemental poverty measure

Conspicuous consumption

Quintile ratio

Lorenz curve

Gini coefficient

Perfect income equality

Perfect income inequality

Gross domestic product (GDP)

National Income and Product Accounts (NIPA)

Flow variable

Production possibilities frontier (PPF)

Intermediate goods

Multiple counting

Government transfer payments

Private transfer payments

Income-expenditure identity

Expenditures approach

Personal consumption expenditures

Durable goods

Nondurable goods

Services

Gross private domestic investment

Final capital goods

Residential fixed investment

Business fixed investment

Inventory investment

Private capital stock

Depreciation

Stock variable

Government purchases

Net exports

Trade balance

Trade deficit

Trade surplus

Balanced trade

Income approach

Compensation

Rental income

Interest income

Profit income

Unincorporated businesses

Sole proprietorships

Partnerships

Unlimited liability

Corporate profits

Limited liability

Dividends

Shareholders

Retained earnings

Taxes on production and imports

National income

Net foreign factor income

Gross National Product (GNP)

Statistical discrepancy

Net Domestic Product (NDP)

Nominal Gross Domestic Product (GDP)

Real Gross Domestic Product (GDP)

Economic growth

Business cycle

Recessions

Expansions

Peak

Trough

Overshooting

Cobweb model

Real GDP growth rate

Per capita real GDP

Real GDP per worker

Imputed value

Extra gross wages

Gross National Happiness (GNH)

Civilian non-institutional population

Not in the civilian non-institutional population

Labor force

Not in the labor force

Employed

Unemployed

Discouraged workers

Labor force participation rate

Unemployment rate

Special unemployment rate

Natural unemployment

Structural unemployment

Frictional unemployment

Cyclical unemployment

Full employment

Potential GDP

Price index

GDP deflator

Inflation rate

Deflation

Substitution bias

New goods bias

Quality change bias

Real interest rate

Nominal interest rate

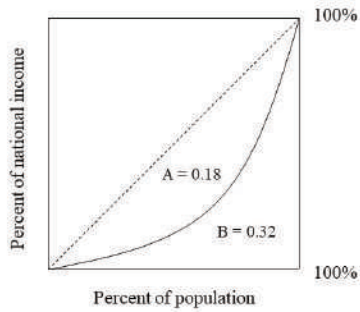
Hyperinflation

Problems for Review

1. In 2017, the poverty threshold for a family of four people with two children under the age of 18 years of age was $24,858. If a family of this size has an income of $23,500, then calculate the income surplus or deficit and the income to poverty ratio. What do these measures allow you to conclude about the poverty status of the family?

2. Suppose the Lorenz curve representing the income distribution in a country looks like the one shown in Figure 12.19.

Figure 12.19: Problems for Review #2



Calculate the Gini coefficient. If the Lorenz curve then shifts and becomes closer to the diagonal line, what should happen to the Gini coefficient and what would you conclude about the degree of income inequality in this country?

3. Use the information below to answer the questions:

- Households spend $2,800 on durable goods.
- The federal government spends $5000 on final services.
- U.S. employee compensation is $11,400.
- State and local governments spend $1,400 on final durable goods.
- New homes are constructed and sold for $1,500.
- Businesses invest in new durable capital goods, spending $1,800.

- U.S. rental income is $800.
- Businesses add $600 worth of goods to their inventories.
- Businesses sell $400 worth of goods from their inventories.
- U.S. interest income is $1,200.
- U.S. businesses sell $900 of durable goods to the rest of the world.
- U.S. profit income is $4,000.
- U.S. businesses purchase $800 of durable goods from the rest of the world.
- U.S. taxes on production and imports are $2,500.
- Households spend $1,500 on nondurable goods.
- Households spend $6,000 on services.
- New business construction amounts to $900.
- Depreciation of the capital stock is $1,200.
- Foreign citizens and businesses operating in the U.S. earn $700.
- American citizens and businesses operating abroad earn $600.

a. What is consumer spending?
b. What is investment spending?
c. What is government spending?
d. What are exports?
e. What are imports?
f. What are net exports?
g. Does a trade deficit or a trade surplus exist? How do you know?
h. What is GDP using the expenditures approach?

i. What is NDP?
j. What is GNP?
k. What is net foreign factor income?
l. What is GDP using the income approach?

4. Use the information in Table 12.11 to calculate nominal GDP, real GDP, and the real GDP growth rate.

Table 12.11: Problems for Review #4 and #5

Year	Apples	Price of apples	Oranges	Price of oranges	Nominal GDP	Real GDP (2004 prices)	Real GDP growth rate	GDP deflator	Inflation rate
2003	250	$0.55	180	$1.10			--		--
2004	255	$0.59	185	$1.21					
2005	262	$0.62	188	$1.27					

5. Complete the rest of Table 12.11. Calculate the GDP deflator and the inflation rate.

6. Suppose the civilian non-institutional population is 275, the number of employed is 152, and the number of unemployed is 9.6. Calculate the unemployment rate. Calculate the labor force participation rate.

7. Use the information in Table 12.12 to calculate the CPI and the inflation rate.

Table 12.12: Problem for Review #7

Year	Units of Food	Price of Food	Units of Fuel	Price of Fuel	Price in the base year	Price in the current year	CPI	Inflation Rate (%)
2007		14		11				--
2008	650	21	780	9				
2009		19		14				
2010		26		18				
2011		20		13				

Assume the base year is 2008.

8. If your nominal income is $2,000 per month in 2018 and the CPI is 1.40 with 2011 as the base year, then what is the real value of your monthly income in constant 2011 dollars?

9. If the nominal interest rate is 9% and the inflation rate is 6.2%, then what is the real interest rate?

Notes

1. Center for Poverty Research, University of California, Davis. “How is Poverty Measured in the United States?” Web. Accessed on April 13, 2018. https://poverty.ucdavis.edu/faq/how-poverty-measured-united-states
2. Ibid.
3. United States Census Bureau. “How the Census Bureau

Measures Poverty." Web. Accessed on April 13, 2018. https://www.census.gov/topics/income-poverty/poverty/guidance/poverty-measures.html

4. Ibid.
5. Institute for Research on Poverty. "How is Poverty Measured in the United States?" University of Wisconsin-Madison. 2016. Web. Accessed on April 13, 2018. https://www.irp.wisc.edu/faqs/faq2.htm
6. United States Census Bureau. "How the Census Bureau Measures Poverty." Web. Accessed on April 13, 2018. https://www.census.gov/topics/income-poverty/poverty/guidance/poverty-measures.html
7. Ibid.
8. Institute for Research on Poverty. "How is Poverty Measured in the United States?" University of Wisconsin-Madison. 2016. Web. Accessed on April 13, 2018. https://www.irp.wisc.edu/faqs/faq2.htm
9. United States Census Bureau. "How the Census Bureau Measures Poverty." Web. Accessed on April 13, 2018. https://www.census.gov/topics/income-poverty/poverty/guidance/poverty-measures.html
10. Blank, Rebecca M. "How We Measure Poverty." Brookings. September 15, 2008. Web. Accessed on April 13, 2018. https://www.brookings.edu/opinions/how-we-measure-poverty/
11. Ibid.
12. Ibid.
13. Center for Poverty Research, University of California, Davis. "How is Poverty Measured in the United States?" Web. Accessed on April 13, 2018. https://poverty.ucdavis.edu/faq/how-poverty-measured-united-states

14. Institute for Research on Poverty. “How is Poverty Measured in the United States?” University of Wisconsin-Madison. 2016. Web. Accessed on April 13, 2018. https://www.irp.wisc.edu/faqs/faq2.htm
15. Center for Poverty Research, University of California, Davis. “How is Poverty Measured in the United States?” Web. Accessed on April 13, 2018. https://poverty.ucdavis.edu/faq/how-poverty-measured-united-states
16. Savage, Michael. “Richest 1% on Target to Own Two-Thirds of All Wealth by 2030.” *The Guardian*. U.S. Edition. April 7, 2018. Web. Accessed on April 14, 2018. https://www.theguardian.com/business/2018/apr/07/global-inequality-tipping-point-2030
17. McConnell and Brue (2008), p. 109, emphasize the relationship between gross investment, depreciation, and net investment when discussing the factors that influence the private capital stock.
18. U.S. Bureau of Economic Analysis, Real Gross Domestic Product [GDPCA], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/GDPCA, April 15, 2018.
19. Prof. Leo Navin first introduced me to a similar version of this model when I was a student in his macroeconomics principles class at Bowling Green State University in 1997.
20. Aslaksen (1999), p. 413.
21. Ibid. Pp. 413
22. Ibid. Pp. 414
23. Ibid. Pp. 415
24. Ibid. Pp. 415
25. Ibid. Pp. 415
26. Ibid. Pp. 414

27. Ibid. Pp. 414
28. Ibid. Pp. 415
29. Ibid. Pp. 415
30. Schultz, Kai. "In Bhutan, Happiness Index as Gauge for Social Ills." *The New York Times*. January 17, 2017. Web. Accessed on April 10, 2018. https://www.nytimes.com/2017/01/17/world/asia/bhutan-gross-national-happiness-indicator-.html
31. Kelly, Annie. "Gross national happiness in Bhutan: the big idea from a tiny state that could change the world." *The Guardian*. U.S. Edition. December 1, 2012. Web. Accessed on April 10, 2018. https://www.theguardian.com/world/2012/dec/01/bhutan-wealth-happiness-counts
32. Ibid.
33. Ibid.
34. Ibid.
35. Schultz, Kai. "In Bhutan, Happiness Index as Gauge for Social Ills." *The New York Times*. January 17, 2017. Web. Accessed on April 10, 2018. https://www.nytimes.com/2017/01/17/world/asia/bhutan-gross-national-happiness-indicator-.html
36. Ibid.
37. Collinson, Patrick. "Finland is the happiest country in the world, says UN report." *The Guardian*. U.S. edition. March 14, 2018. Web. Accessed on April 10, 2018. https://www.theguardian.com/world/2018/mar/14/finland-happiest-country-world-un-report
38. Source: Bureau of Labor Statistics. U.S. Department of Labor. Labor Force Statistics from the Current Population Survey. The figures are seasonally adjusted. Data extracted on April 15, 2018. Only the total population figure was taken from the U.S. Census Bureau. Accessed on April 15, 2018.

39. Source: U.S. Bureau of Labor Statistics. Consumer Price Index. Frequently Asked Questions (FAQS).
40. These sources of bias are commonly cited in mainstream textbooks. See Hubbard and O'Brien (2019), p. 681-682, and Bade and Parkin (2013), p. 592-593.
41. McConnell and Brue (2008), p. 138, consider a similar example when describing the impact of inflation on real income.
42. See Mishkin (2006), p. 80, footnote 10, for a more exact expression.
43. Leubsdorf, Ben. "U.S. GDP Growth Revised Up to 2.9% Rate in Fourth Quarter." *The Wall Street Journal*. Updated March 28, 2018. Web. Accessed on April 15, 2018.

CHAPTER 13

THE THEORY OF EFFECTIVE DEMAND AND THE NEOCLASSICAL SYNTHESIS MODEL

Goals and Objectives:

In this chapter, we will do the following:

1. *Define* Say's Law of Markets and its role in classical economic theory
2. *Describe* how John Maynard Keynes created a revolution in macroeconomic theory
3. *Analyze* the Keynesian Cross model and the Keynesian multiplier effect
4. *Build* the aggregate demand/aggregate supply (AD/AS) model to explain the price level and level of aggregate output
5. *Apply* the AD/AS model to several historical cases from U.S. economic history
6. *Identify* the Neoclassical Synthesis Model and the Post-Keynesian critique of it

This chapter introduces the reader to macroeconomic theory. One major goal is to describe how the key macroeconomic variables from the previous chapter are determined using theoretical models. To really understand these theoretical models, however, it is helpful to explain how the field of macroeconomics developed in the first place. This chapter thus explains the way in which John Maynard Keynes led a revolution in economic thought in the 1930s. Keynes laid the groundwork for what became known as macroeconomics. To understand Keynes's role in the history of economic thought, it is necessary to understand his critique of the classical theory of the aggregate economy and one of its key elements known as Say's Law of Markets. Once Keynes's role is made clear, it will be easier to understand his theory of output and employment. This theory is represented in the Keynesian Cross model and Keynes's concept of the spending multiplier. We will also build the aggregate demand/aggregate supply (AD/AS) model so as to provide an explanation, not only for aggregate output and employment, but also for the general level of prices. Using the AD/AS model, we will use it to understand actual changes in aggregate output and the price level that occurred during different periods in U.S. economic history.

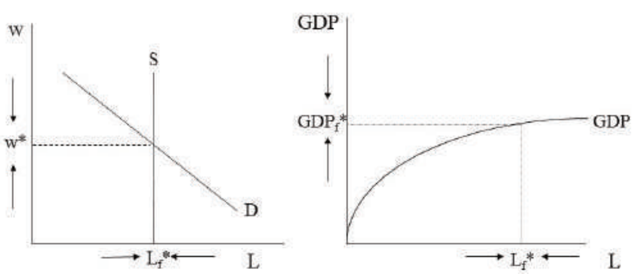
Say's Law and the Classical Theory

It might strike the reader as strange that economics is divided into two separate fields referred to separately as "microeconomics" and "macroeconomics." This division has not always existed. During the eighteenth and early nineteenth centuries, the discipline was simply referred to as "political economy." Economists generally regarded questions at the individual level and questions at the societal level as questions to be treated together. During the marginalist revolution of the late nineteenth century, however, some economists became very much preoccupied with questions of efficiency and individual optimization. Larger questions having to do with economic growth, the general level of prices, and overall employment became increasingly separate from this specialized study of atomistic behavior. In 1936, this bifurcation intensified when John Maynard Keynes published his famous book titled, *The General Theory of Employment, Interest, and Money*. Published shortly after the Great Depression, Keynes offered a theory to explain how aggregate output, employment, and prices are determined. His explanation sharply contrasted with early neoclassical microeconomic theories of how individual markets function.

In his critique of early neoclassical microeconomic theory, Keynes referred to the theory as the "classical" theory. Keynes's decision to do so stemmed in part from the fact that early neoclassical theory had a laissez-faire orientation that was very similar to the laissez-faire orientation of the classical theories of Adam Smith and David Ricardo. His decision also stemmed from his critique of one aspect of classical economics that was central to early neoclassical thinking, referred to as **Say's Law of Markets**. According to Say's Law of Markets (or Say's Law, for short), every supply creates its own demand. That is, if a commodity is produced, it will generate enough factor income for the owners of land, labor, and capital to purchase the produced commodity. After all, the price of the commodity may be divided into its component parts of rent, wages, and profits. Hence, these incomes will be sufficient to realize the price of the commodity. The stunning implication of this simple argument at the level of the aggregate economy is that enough income will always exist to purchase the entire output of commodities. Therefore, no general gluts or periods of overproduction are possible. Using nothing but logic, one can conclude that major depressions should never occur. If they do occur, they must be caused by some interference with the free flow of commodities and money, and government interference is a likely culprit. Say's Law thus supported the laissez-faire orientation of classical economics and early neoclassical economics later.

Another way to understand the classical theory of output and employment is to consider the way in which competition in the labor market will lead to a full employment equilibrium outcome, as depicted in Figure 13.1.[1]

Figure 13.1: The Classical Theory of Employment and Output

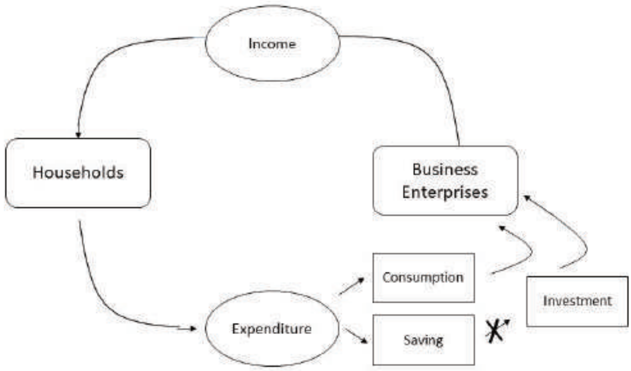


If the market wage (w) is above the equilibrium wage (w*), then a surplus of labor or unemployment exists. Competition will force the market wage down until the market wage coincides with the equilibrium wage. The surplus will be eliminated and the quantity of labor demanded will equal the available labor force (L_f*). With the economy operating at full employment, the economy will be able to produce at the full employment level of GDP (GDP_f*). The full employment GDP is shown in Figure 13.1 using the **aggregate production function**. The aggregate production function exhibits diminishing returns to labor, which assumes that all other factors of production (e.g., land and capital) are held constant. Of course, all available land and capital will be fully employed as well, assuming that the markets for these resources have cleared as well. Given the available production technology and the fully employed resources, the economy will produce the potential GDP.

John Maynard Keynes's *General Theory of Employment, Interest, and Money*

John Maynard Keynes was well-trained in neoclassical economic theory. In fact, Keynes's critique of neoclassical theory was not intended to undermine it entirely. Instead, Keynes regarded that theory as applicable to a special case only, namely the case of an economy operating at full employment. That is, once the economy operates at full employment, then all of the efficiency conclusions of neoclassical theory apply once again. Keynes's theory was intended to be a more general theory of how the aggregate economy functions, however, and so it offered a framework for thinking about periods during which the economy failed to achieve full employment. To make this argument, Keynes had to attack that central tenet of classical economic theory known as Say's Law. The simple flow diagram in Figure 13.2 helps to illustrate Keynes's argument.[2]

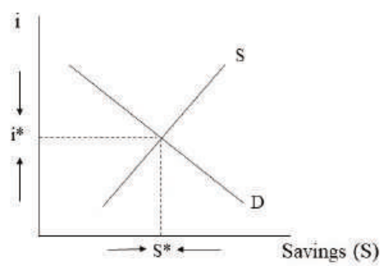
Figure 13.2: Keynes's Objection to Say's Law of Markets



As Figure 13.2 shows, business enterprises make payments to households in exchange for the resources households sell to businesses. Households receive factor income for the land, labor, and rent that businesses purchase. The businesses use these factors of production to produce finished commodities. The households then spend this factor income on the finished commodities, but it is here that we deviate from our earlier discussion of the neoclassical circular flow model from Chapter 3. Obviously, households will choose to save some of their income and not spend it. The classical economists and early neoclassical economists understood this point well. They argued that the saving that occurred would flow into the loanable funds market where it would be transferred to individuals and firms who would invest it. These investments in new capital goods would add to the demand for consumer goods and ensure that all of the finished commodities were sold. Of course, these borrowers would pay interest to the lenders for the use of the borrowed funds, but the circular flow would be maintained and Say's Law would hold.

Keynes rejected this version of the circular flow on the grounds that the saving that occurred would not all be transformed into the investment demand for capital goods. For Keynes, the failure of all savings to become investment had different possible sources. Without going into the details, the primary reasons had to do with the rate of interest.[3] Consider the loanable funds market as shown in Figure 13.3.

Figure 13.3: The Loanable Funds Market and the Equilibrium Rate of Interest



If a surplus of savings exists, then in the classical theory, the market rate of interest (i) will fall to the equilibrium level of i* and clear the loanable funds market. Aggregate saving will equal aggregate investment. If, however, other forces are acting on the rate of interest that prevent it from falling, then the surplus of savings will persist and not all savings will be invested. As a result, not all finished commodities will be sold due to insufficient investment demand. Alternatively, it might be that the market clearing level of the rate of interest (i*) does not occur at a market rate of interest that is greater than or equal to zero due to a very large supply of savings and a relatively low level of investment demand. In that case, the market rate of interest will not be able to fall enough to clear the market for loanable funds and the surplus of savings will persist. In this case as well, an excess supply of finished commodities or a glut will occur. The result will be falling production as firms scale back production, falling employment as they lay off workers, and falling prices due to the excess supplies of commodities. All these results were observed during the Great Depression, which is exactly what Keynes developed his theory of effective demand to explain with the hope of ending depressions by means of enlightened government economic policy.

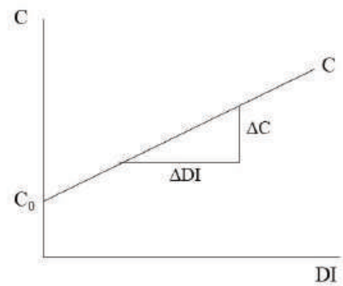
The Consumption Function and the Saving Function

To understand the theory that Keynes developed, it is necessary to begin with the construction of some foundational elements. We first introduce the **consumption function** and show how it is represented graphically. The consumption function suggests that a positive relationship exists between the level of current consumption expenditures that households are planning and the current level of disposable income. That is, as households acquire more disposable income, their level of consumption rises, other factors held constant (ceteris paribus). Planned consumption thus depends on the disposable income of households. That is, C is a function of DI or C = f (DI). Alternatively, C is the dependent variable and DI is the independent variable, according to this theory. This relationship can be represented mathematically as follows:

$$C = C_0 + \frac{\Delta C}{\Delta DI} DI$$

In the consumption function, C_O represents **autonomous consumption**. Autonomous consumption is the level of consumption expenditures that households choose even if their DI falls to zero (i.e., an amount that is independent of income). Such expenditures might be possible if households rely on their savings to finance consumer expenditures. Borrowing would be another way in which consumer expenditures might be positive even when DI equals zero. The consumption function also includes ΔC/ΔDI, which represents the **marginal propensity to consume** (mpc). The mpc refers to the additional consumption expenditures that households choose for each additional dollar of disposable income received. Figure 13.4 reveals that the level of autonomous consumption determines the vertical intercept of the consumption function in the graph. Similarly, the mpc represents the slope of the consumption function in the graph.

Figure 13.4: The Consumption Function



Because the mpc is assumed to be fixed at every level of DI, the slope is constant and the consumption curve graphs as a straight line. To consider a simple example, suppose that the consumption function is C = 200+0.75DI. In this case, even if DI is equal to zero, the households will spend $200 billion. Also, for every $1 of additional DI that the households receive, they will consume an additional $0.75.

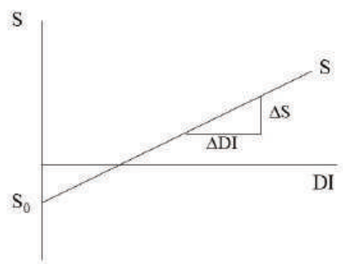
It is worth noting that only a change in DI can cause a movement along the consumption curve in the graph whereas a change in autonomous consumption will shift the consumption curve up or down. Various factors can shift the consumption curve up or down. One example is a change in household wealth, which is distinct from disposable income. The reader might recall that DI is a flow variable. That is, it is measured per period of time (e.g., per year). Household wealth, on the other hand, is a stock variable and is measured at a point in time. If households experience a reduction in household wealth due to a recession that causes asset values (e.g., stock prices) to fall, then the consumption expenditures will fall at every level of disposable income and the consumption curve will shift in a downward direction. Alternatively, an economic boom that raises household wealth will cause the consumption curve to rise at every income level and will lead to an upward shift.

Just like we can represent the level of planned household consumption expenditures, we can also represent the level of planned household saving as the level of disposable income changes. The **saving function** suggests that a positive relationship exists between planned household saving and disposable income. That is, as households acquire more disposable income, their level of saving rises, other factors held constant (ceteris paribus). Planned saving thus depends on the disposable income of households. That is, S is a function of DI or S = f (DI). Alternatively, S is the dependent variable and DI is the independent variable, according to this theory. This relationship can be represented mathematically as follows:

$$S = S_0 + \frac{\Delta S}{\Delta DI} DI$$

In the saving function, S_O represents **autonomous saving**. Autonomous saving is the level of saving that households choose even if their DI falls to zero (i.e., an amount that is independent of income). Such "saving" might occur if saving is negative. That is, households do not save but actually borrow or draw down past savings. The saving function also includes ΔS/ΔDI, which represents the **marginal propensity to save** (mps). The mps refers to the additional saving that households choose for each additional dollar of disposable income received. Figure 13.5 reveals that the level of autonomous saving determines the vertical intercept of the saving function in the graph. Similarly, the mps represents the slope of the saving function in the graph.

Figure 13.5: The Saving Function

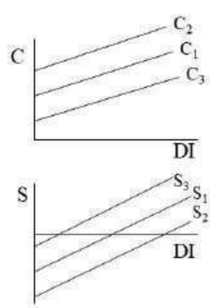


Because the mps is assumed to be fixed at every level of DI, the slope is constant and the saving curve graphs as a straight line. To consider a simple example, suppose that the saving function is S = -200+0.25DI. In this case, even if DI is equal to zero, the households will save -$200 billion. Also, for every $1 of additional DI that the households receive, they will save an additional $0.25.

It is worth noting that only a change in DI can cause a movement along the saving curve in the graph whereas a change in autonomous saving will shift the saving curve up or down. Various factors can shift the saving curve up or down. A rise in household wealth will lead to a fall in saving (as consumption rises) and shift the saving curve downward. A fall in household wealth will lead to a rise in saving and shift the saving curve upward. As in the case of consumption, planned saving by households represents a flow variable.

It is worth taking a moment to reflect on the relationship between the vertical intercepts of the consumption and saving functions. In both cases, the level of DI = 0, which implies that C+S = 0. This equation further implies that when DI = 0, C = -S. In other words, the vertical intercept of the saving curve is the negative of the vertical intercept of the consumption curve. Therefore, whenever the consumption curve shifts upward, the saving curve will shift downwards, and vice versa. Figure 13.6 shows how a shift of the consumption curve will lead to an opposite change in the saving curve.

Figure 13.6: Determinants of Consumption: Relaxing the Ceteris Paribus Assumption



One should also consider the relationship between the

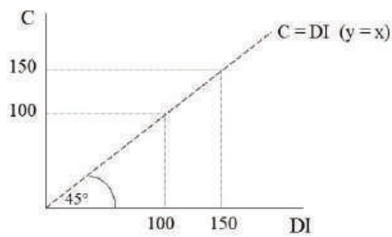
mpc and the mps. Suppose we add the two measures together as follows:
$mpc + mps = \frac{\Delta C}{\Delta DI} + \frac{\Delta S}{\Delta DI} = \frac{\Delta C + \Delta S}{\Delta DI} = \frac{\Delta DI}{\Delta DI} = 1$
In other words, the sum of the mpc and the mps is always equal to 1. This result is very intuitive. If the households receive $1 of additional DI and they consume $0.75 of it, then the remaining $0.25 must be saved.

Now that we understand the essential building blocks of Keynes's theory, we need to consider a national income accounts identity that relates disposable income (DI) to consumption (C) and saving (S).

$DI = C + S$

We then consider the case in which S = 0 and thus C = DI. To represent this case graphically, we introduce a **reference line** that has a 45 degree angle relative to the horizontal axis as shown in Figure 13.7.

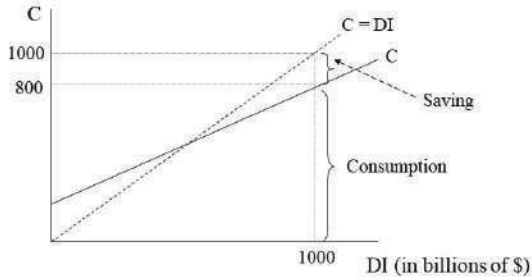
Figure 13.7: The Building Blocks of Keynes's Macroeconomic Theory



As Figure 13.7 indicates, at each level of DI, the level of C that corresponds to it is found on the reference line. That is, the horizontal distance to that level of DI will be exactly the same as the vertical distance from that level of DI up to the 45 degree line. Another way to think about this reference line is that it has a slope equal to 1 and a vertical intercept equal to zero, much like the equation y = x in the case that y is placed on the vertical axis and x is placed on the horizontal axis.

Of course, saving is not expected to equal zero at every level of DI and so this case is not a realistic one. When we place the reference line on a graph with the planned consumption curve as in Figure 13.8, however, we see that the difference between the two lines carries a special meaning.

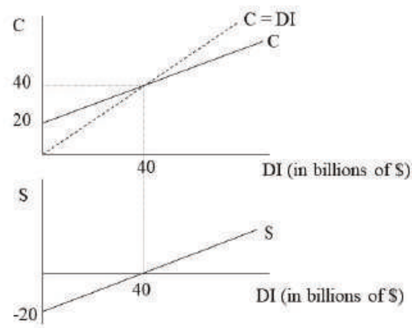
Figure 13.8: The Consumption Function and the Reference Line



As Figure 13.8 shows, at a level of DI of $1000 billion, the reference line allows us to represent that amount vertically. It is then much easier to see which part of the DI is spent on consumption and which part remains for saving. In this case, $800 billion is consumed and $200 billion is saved. In general, at every level of DI at which the C line is below the reference line, positive saving exists. The reader will notice that the lines intersect at a single point. At this point, planned consumer spending equals DI, which is true for all points on the reference line. When planned consumer spending equals DI, we refer to this income level as the break-even income level. In other words, households have just enough DI to finance their consumption expenditures, no more and no less. Finally, if DI is at a low enough level that the C line is above the reference line, then planned household consumption expenditures exceed DI and the households must be borrowing or relying on past savings to finance the excess consumption.

We can also consider how the graph of the consumption function relates to the graph of the saving function. Figure 13.9 depicts the relationship between the two graphs.

Figure 13.9: The Saving Function and the Break-Even Income Level



At the break-even income level of $40 billion in the top graph, C = DI and thus S = 0. Therefore, the bottom graph shows the saving function crossing the DI axis at $40 billion, indicating a level of saving equal to zero at that level of DI. As we move to the right of the break-even income level in the top graph, we see that saving becomes positive and continues to grow as the gap between the two lines becomes larger. Therefore, in the bottom graph, the level of saving becomes positive and continues to grow when DI rises above $40 billion.

Let's suppose we are given the following consumption function:

$C = 20 + 0.75DI$

It turns out that it is possible to derive the saving function from this information. Recall that DI = C+S. Substituting the consumption function into this equation yields DI = 20+0.75DI+S. Solving for S yields S = DI – 0.75DI – 20, which may be simplified as follows:

$S = -20 + 0.25DI$

It should be clear that we can use a shortcut method to obtain the saving function from the consumption function. If we begin with the consumption function and negate the vertical intercept of 20, then we obtain the vertical intercept of -20 for the saving function. Furthermore, if we subtract the mpc of 0.75 from 1, then we obtain the mps of 0.25 since the two marginal propensities always add up to 1.

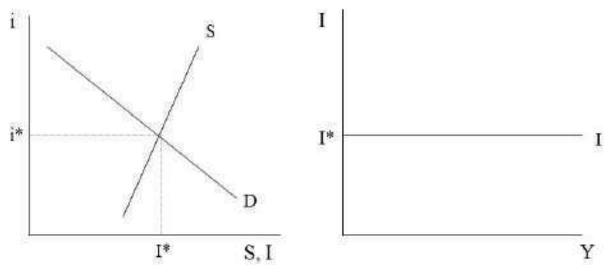
The Keynesian Cross Model for a Private, Closed Economy

It should be clear that Keynes radically departed from the early neoclassical economic theory in which he was trained. In Keynes's theory, aggregates, like households, business enterprises, and the government, take center stage. Individual economic agents do not play an important role in the theory. Also, because households and businesses tend to behave in a collective fashion (e.g., consuming more when DI rises or investing less when business expectations turn sour), mass psychology becomes the primary explanation for these behaviors rather than individual rationality and serves as an alternative conceptual point of entry.[4] For example, households have a propensity to consume so much more when DI rises. Nevertheless, the unidirectional logic of neoclassical theory is preserved in Keynes's theory.[5] That is, one variable affects another in a single causal direction only. For example, a rise in DI causes a rise in consumer spending, but not vice versa.

Based on this theoretical foundation, it is now possible to develop the basic Keynesian theory of output and employment. The **Keynesian Cross model**, which is also sometimes referred to as the **aggregate expenditures model**, uses these theoretical tools to explain the equilibrium levels of aggregate output and employment that emerge. To keep the model simple, we initially assume that the economy is private and closed. That is, only households and businesses exist. Because it is a purely private economy, no government exists. Because it is a closed economy, no foreign trade exists. The price level is also assumed to be constant and so the Keynesian Cross model does not explain prices, only output and employment. Finally, it is assumed that DI is equal to real GDP. In the national income accounts, the income that American households have for consumption and saving (DI) is not equal to real GDP due to the presence of depreciation, net foreign factor income, taxes, and transfer payments. With a closed, private economy without depreciation, no such adjustments need to be made, and DI is equal to real GDP.

To build the Keynesian Cross model, we need to explain the determination of planned investment spending. As explained previously, the level of planned investment spending is determined in the loanable funds market. Savers lend funds to borrowers at interest, and their competitive interaction determines a particular quantity of loanable funds exchanged. These loanable funds are invested in new production plants and equipment. If we assume that the level of planned investment spending is independent of the level of real GDP, then we can represent its determination in the loanable funds market as shown in Figure 13.10 below.

Figure 13.10: The Determination of Planned Investment Spending



As the reader can see in the graph on the right, investment spending (I) is equal to I_O and is thus constant at all levels of real GDP (Y).

It is now possible to write two equations representing the two types of planned expenditures in this simple economy. The consumption function representing the planned consumption of households may be written as:

$$C = C_0 + \frac{\Delta C}{\Delta Y} Y$$

The reader should notice that real GDP (Y) has replaced DI in the consumption function. The reason, of course, is the assumption that real GDP is equal to DI in this economy. The second equation indicates that investment spending (I) is constant, as previously noted.

$$I = I_0$$

If we combine these two types of planned spending, we can obtain a planned aggregate expenditures (A) function as follows:

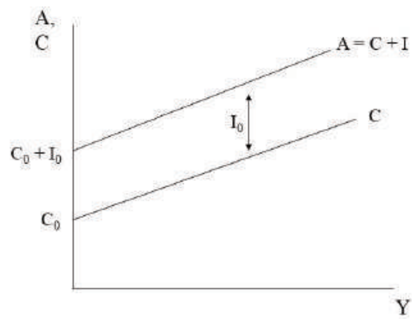
$$A = C + I = C_0 + \frac{\Delta C}{\Delta Y}Y + I_0$$

This planned aggregate expenditures function may be written as follows by rearranging the terms:

$$A = (C_0 + I_0) + \frac{\Delta C}{\Delta Y}Y$$

In this function, $C_O + I_O$ represents autonomous spending. That is, households and businesses will select this level of planned spending regardless of the level of real GDP. The second portion, (ΔC/ΔY)Y, represents **induced spending**. Induced spending is planned aggregate spending that is directly related to the level of real GDP. If we place the planned consumer spending (C) curve and the planned aggregate expenditures (A) curve on the same graph, we obtain the graph shown in Figure 13.11.

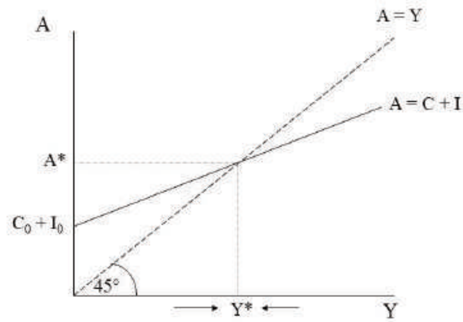
Figure 13.11: The Planned Aggregate Expenditures Curve



As the reader can see, the planned aggregate expenditures curve has a vertical intercept equal to the level of autonomous spending. The vertical distance between A and C is constant and equal to I_O at every level of DI. The slope of the aggregate expenditures curve is also equal to the mpc because the A and C curves are parallel to one another.

Armed with this new tool, we can represent the equilibrium level of real GDP on the graph. Figure 13.12 shows that the equilibrium level of real GDP occurs at the intersection of the aggregate expenditures curve and the 45-degree reference line.

Figure 13.12: The Determination of Equilibrium Real GDP



At this point, planned aggregate spending equals real aggregate output and so all plans are perfectly satisfied by the level of production. The reason this level of real GDP is the equilibrium level can be understood by considering what will occur if the economy is not producing at this point. Suppose that the level of real output is above Y* in the graph. In that case, planned aggregate spending is below the level of real output as shown on the 45-degree line. That is, A < Y. As a result, firms will be producing more than households and firms wish to purchase. As a result, business inventories will build up as a result of the excess supply of commodities. The consequence will be a drop in the level of production as firms cut production. Real GDP will then fall towards the equilibrium level. Conversely, if the level of real GDP is below Y* in the graph, then planned aggregate spending is above the level of real output as show on the 45-degree line. That is, A > Y. As a result, firms will be producing less than households and firms wish to purchase. As a result, business inventories will be depleted as a result of the excess demand for commodities. The consequence will be a rise in the level of production as firms raise production. Real GDP will then rise towards the equilibrium level.

It is also possible to calculate the equilibrium real GDP given a specific consumption function and level of investment. For example, suppose that the following two equations represent an economy:

$C = 200 + 0.75Y$

$I = 100$

Given this information, it is possible to obtain the planned aggregate expenditures function by simply adding the two equations together:

$A = 300 + 0.75Y$

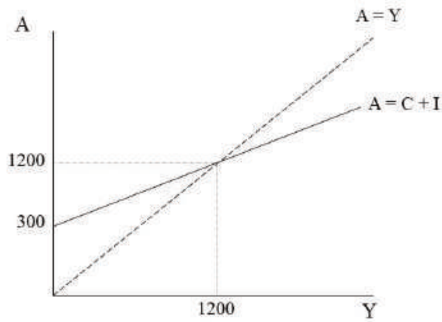
To obtain the equilibrium real GDP for this economy, we need to use the equilibrium condition:

$A = Y$

Plugging in the planned aggregate expenditures function and solving for Y, we obtain the equilibrium level of real GDP.

$300 + 0.75Y = Y$
$Y = 1200$
Figure 13.13 represents the solution graphically.

Figure 13.13: The Calculation of Equilibrium Real GDP in the Keynesian Cross Model



The graph shows that the equilibrium level of real GDP occurs at the intersection of the reference line and the planned aggregate expenditures curve. Solving the two equations simultaneously yields the solution.

It turns out that it is possible to think about the determination of the equilibrium real GDP from another angle. That is, by using the saving function and considering the level of investment spending, it is possible to arrive at a different but related equilibrium condition. To understand this point, consider the equilibrium condition that we have been using up to this point:

$$A = Y$$

The reader should recall that planned aggregate spending (A) is the sum of planned consumer spending (C) and planned investment spending (I). Real GDP or real income (Y) is either consumed (C) or saved (S). If we break down each term in the equation into its component parts, we obtain the following:

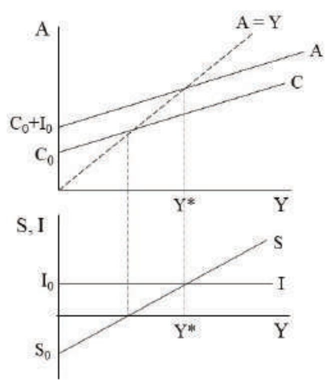
$$C + I = C + S$$

It is easy to see that the level of consumer spending may be subtracted from both sides of this equation to yield a new equilibrium condition:

$$I = S$$

In equilibrium then, planned investment and saving must be equal. Figure 13.14 represents this solution graphically and relates it to the Keynesian Cross model that we have already discussed.

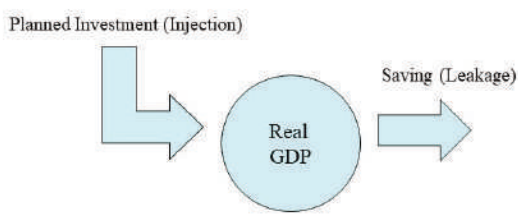
Figure 13.14: The Saving-Investment Approach to the Determination of Equilibrium GDP



Because planned investment is constant at all levels of GDP, it is represented as a horizontal line in the graph. The saving curve, on the other hand, slopes upward for reasons already discussed. The equilibrium level of real GDP occurs at the intersection of the two lines.

The intuition behind the movement to equilibrium in this graph is a little different from the intuition behind the movement to equilibrium in the Keynesian Cross model. In the saving-investment graph, saving may be thought of as a leakage from the spending stream that creates real GDP. Planned investment, on the other hand, is an injection into the spending stream. Planned investment thus raises equilibrium real GDP, and saving reduces equilibrium real GDP. If the two forces exactly balance, then real GDP should remain stable. The manner in which these two forces act on real GDP may be depicted as in Figure 13.15.

Figure 13.15: Injections of Investment and Leakages of Saving



In Figure 13.14, if the level of real GDP is below the equilibrium level of Y*, then planned investment exceeds saving. With the injection larger than the leakage, the result is a rise in real GDP and a movement towards the equilibrium level. On the other hand, if the level of real GDP is above the equilibrium level of Y*, then saving exceeds planned investment. With the leakage larger than the injection, the result is a fall in real GDP and a movement towards the equilibrium level.

It is also possible to arrive at the answer algebraically using the same information we used when discussing the Keynesian Cross model. Because we know the shortcut method of deriving the saving function from the consumption function, we can write the saving function alongside the consumption function as follows:

$$C = 200 + 0.75Y$$
$$S = -200 + 0.25Y$$

Given that I = 100, we use the new equilibrium condition as follows:

$$I = S$$

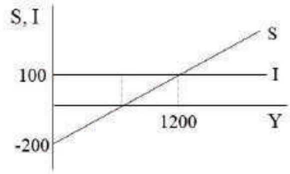
$$100 = -200 + 0.25Y$$

We then solve for Y to obtain:

$$Y = 1200$$

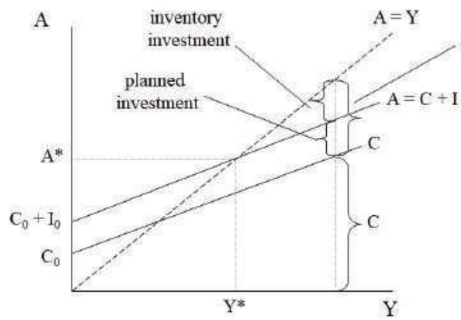
The reader will note that this equilibrium real GDP is the same as the one calculated in the Keynesian Cross model. Figure 13.16 shows the solution in a saving-investment graph.

Figure 13.16: The Calculation of Real GDP in the Saving-Investment Graph



The reader might be a bit confused by the claim that saving equals investment only in equilibrium. After all, it was argued in Chapter 12 that GDP is always equal to the sum of consumer spending, investment spending, government spending, and net export spending. In a closed, private economy without government spending or net exports, GDP should always equal consumer spending plus investment spending. Therefore, if GDP is the sum of consumer spending and saving, then saving should always equal investment spending. The reason for the apparent contradiction is that the equilibrium condition only refers to **planned investment** whereas the measurement of GDP includes **actual investment**. In other words, saving will always equal actual investment, which includes both planned investment spending and inventory (unplanned) investment spending. At the same time, saving may not equal planned investment spending. Figure 13.17 clarifies this point.

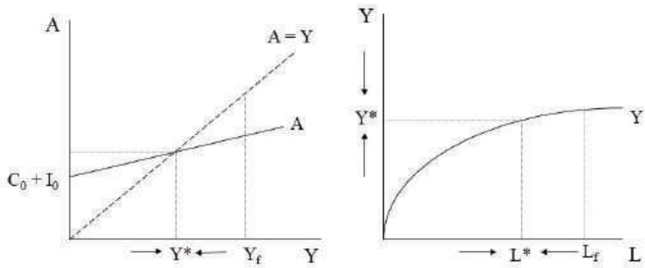
Figure 13.17: A National Income Accounts Identity:
Saving ≡ Actual Investment



As the reader can observe from Figure 13.17, if GDP exceeds the equilibrium level of Y*, then saving exceeds planned investment. At the same time, saving is still equal to actual investment because the part of what is saved that is not intentionally invested is unintentionally invested in inventories. That is, not all commodities are sold and so firms will add these commodities to their inventories, which counts as unplanned inventory investment. Therefore, the national income accounts identity holds, but the equilibrium condition, which requires that saving equals planned investment, does not hold.

It is now possible to clearly distinguish the Keynesian model of output and employment from the classical model. Figure 13.18 shows clearly that the level of planned aggregate expenditure determines the equilibrium level of GDP, which is likely to be below the full employment GDP (Y_f) at a point in time.

Figure 13.18: The Keynesian Theory of Employment and Output



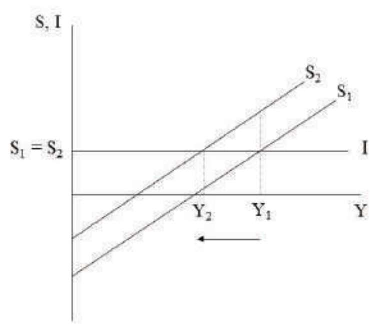
Furthermore, as businesses adjust production in the movement to equilibrium, they also adjust their workforces. Employment, therefore, moves towards an equilibrium level of L* that corresponds to the equilibrium level of real GDP as shown on the production function. Employment also ends up below the full employment level of L_f. Hence, Keynes's theory is one of unemployment equilibrium, and it reveals the case of full employment GDP to be a special case. That is, aggregate planned spending would need to be just high enough to produce the full employment GDP as the equilibrium GDP. Keynes could, therefore, argue that his theory was a more general theory than the classical theory of employment and output.

The Paradox of Thrift

An interesting application of the model allows us to understand what macroeconomists mean by the **paradox**

of thrift. In Chapter 2 we learned that an increase in saving leads to greater capital accumulation and an expansion of production possibilities. When considering short run macroeconomic fluctuations, however, saving may lead to a very different result. Suppose that all households decide to increase saving at every level of GDP. In other words, autonomous saving rises. When this change occurs, the saving curve shifts upward as shown in Figure 13.19.

Figure 13.19: Application: The Paradox of Thrift



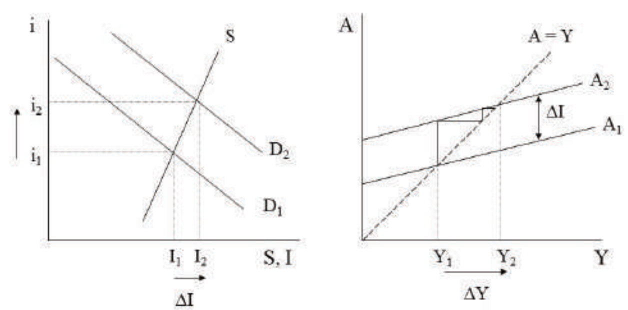
Initially, saving rises at Y_1, but this increase in saving causes a discrepancy between saving and planned investment spending. Specifically, saving rises above planned investment spending. With the leakage of saving being higher than the injection of planned investment, real GDP begins to fall. The drop in real GDP causes a movement along the new saving curve. In other words, **induced saving** declines. The reduction in saving continues until it once again equals planned investment and the economy returns to equilibrium at Y_2. The problem for the economy is that the high level of saving has led to a recession (i.e., falling real GDP) rather than to economic growth, as suggested by the production possibilities model. Furthermore, and rather paradoxically, even though all households decided to save more, aggregate saving returns to the same level that previously existed.[6] The reason is that the recession has led to falling incomes, which reduces saving.

This example illustrates what neoclassical economists refer to as the **fallacy of composition**. People often assume that when one person acts in a particular way and achieves it that the same will hold true for groups of people. To use a classic example, an individual attending a sporting event might stand up to obtain a better view of the field. This strategy works, but if everyone has the same idea, then when all stand, no one has a better view than before. Similarly, one household might save more of its income successfully, but if all households do the same, then no one is able to save more.

The Multiplier Effect

Now that the Keynesian Cross model has been developed, we can consider one of Keynes's most important contributions to our understanding of the manner in which changes in spending affect the overall economy. The Keynesian **multiplier effect** refers to the way in which an increase in a specific component of spending, such as investment spending, raises real GDP by a multiple of the spending increase. To illustrate this point, we will consider the most volatile component of aggregate spending and the impact of changes in it on real GDP. Investment spending tends to be highly volatile for a number of reasons. It is influenced by business expectations about future profits, interest rates, technological change, and taxes on profit income. Suppose that business expectations about future profitability improve significantly. With businesses feeling more optimistic about the future of the economy, a collective rush to invest in new capital takes place. Keynes referred to such impulses during periods of business optimism as **animal spirits**. The consequence is a rise in demand in the loanable funds market. Such an increase leads to a rise in the equilibrium level of loanable funds exchanged and to a rise in planned investment spending as shown in Figure 13.20.

Figure 13.20: The Investment Multiplier in the Aggregate Expenditures Model



The rise in planned investment raises the aggregate expenditures curve as shown in Figure 13.20. The result is a higher level of equilibrium real GDP and the economy experiences an economic boom. Interestingly, the graph on the right in Figure 13.20 suggests that the level of real

GDP rises by more than the rise in planned investment. Because the slope of the reference line is equal to 1 and the difference between the old and new A curves is equal to the change in planned investment, the first move up along the reference line would indicate a rise in real GDP equal to the rise in planned investment spending. As the reader can see in the graph, however, real GDP rises by more than this amount. Hence, a multiplier effect is implicit in the Keynesian Cross model.

It is worth asking why this change occurs. Intuitively, when businesses engage in new investment spending, they raise real GDP by the amount of the investment spending. This spending is received as income by the households though. Once received, the households spend a portion of the additional income, which is determined by the marginal propensity to consume. That additional expenditure is received by the households as well, and part of it is spent. This cycle continues indefinitely but the spending that occurs in the successive rounds becomes smaller and smaller due to the saving that occurs in each round. Ultimately, aggregate real GDP rises by a finite amount but also by a multiple of the initial amount of investment spending.

It is possible to derive a formula that tells us the exact impact that a change in investment spending has on real GDP, other factors held constant. Let's define the investment multiplier as the ratio of the change in real GDP to the change in planned investment spending. By taking into account the way in which the households engage in an infinite series of consumer spending rounds, we can prove that the multiplier is positively related to the marginal propensity to consume as shown below.

$$\frac{\Delta GDP}{\Delta I} = \frac{1}{1-mpc} = \frac{1}{mps}$$

Figure 13A.1 provides the details of the proof for any interested readers. Our main purpose, however, is to understand the intuition behind the formula and to learn how to use it to calculate changes in real GDP. To understand how to use the formula, let's assume that planned investment spending rises by 100. If the mpc is 0.75, then we can calculate the investment multiplier by plugging the mpc into the formula as follows:

$$\frac{\Delta GDP}{\Delta I} = \frac{1}{1-mpc} = \frac{1}{1-0.75} = \frac{1}{0.25} = 4$$

Because investment spending rises by 100, we can write the equation as follows:

$$\frac{\Delta GDP}{100} = 4$$

By solving for the change in GDP, it is clear that real GDP will rise by $400 billion in this case. Alternatively, the multiplier implies that for every $1 of additional investment spending, real GDP will rise by $4.

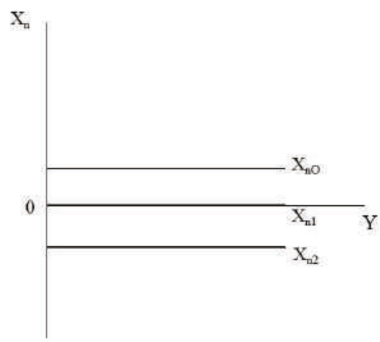
The Keynesian Cross Model of an Open, Mixed Economy

In this section, we would like to expand the Keynesian Cross model to include international trade and government activity. To account for these types of expenditure, we will make some simplifying assumptions. First, we will assume that net exports are exogenously given as follows:

$$X_n = X_{n0}$$

In other words, net exports are at the same level regardless of the level of real GDP as shown in Figure 13.21.

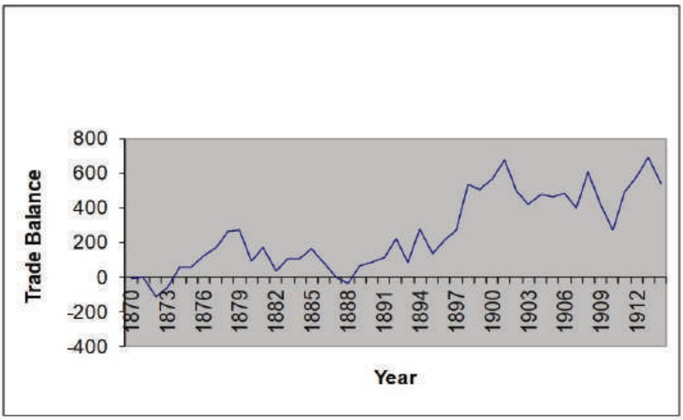
Figure 13.21: The Net Export Curve



As Figure 13.21 shows, net exports may be positive (X_{nO}), equal to zero (X_{n1}), or negative (X_{n2}). The reader should recall that net exports equal exports (X) minus imports (M). As explained in Chapter 12, if net exports are positive, then a **trade surplus** exists (X > M). If net exports are negative, then a **trade deficit** exists (X < M). If net exports are equal to zero, then **balanced trade** exists (X = M).

Although trade deficits have been common in the United States since the mid-1970s, at an earlier time in its history, the U.S. ran trade surpluses. Trade surpluses were common in the U.S. in the late nineteenth and early twentieth centuries as a result of large agricultural surpluses due to high productivity in agriculture. The surplus commodities were exported and made possible a period of sustained trade surpluses as shown in Figure 13.22.[7]

Figure 13.22: The U.S. Balance of Trade from 1870-1914 (in millions of dollars)



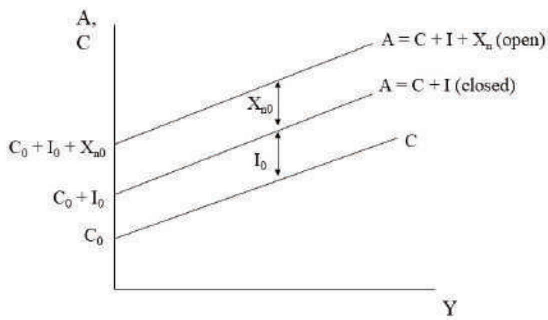
In the Keynesian Cross model, we can now add net exports to the consumption function and level of planned investment spending to obtain the aggregate expenditures function as follows:

$$A = C + I + X_n$$

$$A = (C_0 + I_0 + X_{n0}) + mpc \cdot Y$$

As the reader can see, autonomous spending now includes net export spending. Otherwise, the aggregate expenditures function is the same. Because net exports may be positive, negative, or equal to zero, the vertical intercept may be above, below, or the same as the aggregate expenditures curve for the closed, private economy. Figure 13.23 shows the case of a trade surplus and the impact that it has on the position of the aggregate expenditures curve relative to that of a closed, private economy.

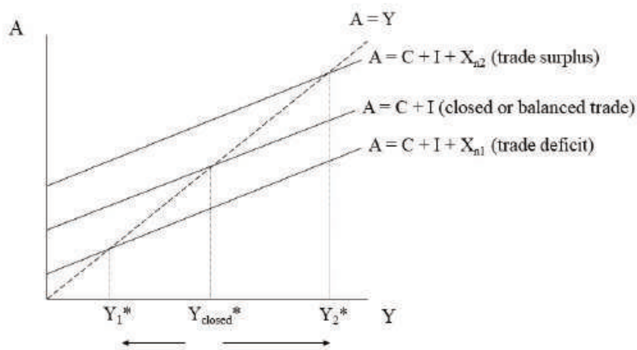
Figure 13.23: Net Exports as a Component of Aggregate Expenditures



As the figure shows, the opening of the economy shifts the aggregate expenditures curve upward due to the trade surplus that results.

The impact of trade on the equilibrium real GDP should also be clear. Figure 13.24 shows what happens when the economy runs a trade surplus or a trade deficit.

Figure 13.24: Equilibrium GDP in an Open, Private Economy



As Figure 13.24 shows, a trade surplus raises the aggregate expenditures curve and increases the equilibrium real GDP. A trade deficit, on the other hand, lowers the aggregate expenditures curve and reduces the equilibrium real GDP. Balanced trade leaves the aggregate expenditures curve unchanged and leaves the equilibrium real GDP unchanged. The case of balanced trade demonstrates that the spending by foreigners on the economy's exports is exactly canceled by the spending of domestic buyers on imports from other countries in terms of the impact on real GDP. This analysis also allows us to draw a conclusion about the desirability of a trade surplus and the disadvantage of a trade deficit. Trade deficits appear to be harmful because they lower the nation's equilibrium real GDP and employment level. Trade surpluses, on the other hand, appear to be beneficial because they raise the nation's aggregate output and employment.

It is important to consider the various factors that lead to trade surpluses and trade deficits.[8] Income levels in other countries certainly play a role. If trading partners undergo economic expansions and incomes are rising, then foreigners will buy more U.S. exports and the trade balance will improve (i.e., net exports will rise). If trading partners experience recessions and incomes are falling, then foreigners will buy fewer U.S. exports and the trade balance will worsen (i.e., net exports will fall).

A second factor affecting the trade balance is tariff policy. A **tariff** is simply a tax on imported commodities. If tariffs are imposed, then prices of imports rise and the quantity of imports will decline. This change should improve the trade balance, possibly causing a trade surplus. At the same time, however, other nations might retaliate by imposing their own tariffs, which might reduce the nation's exports. In that case, the overall impact on net exports appears to be uncertain. Such retaliatory tariffs were imposed during the 1930s after the U.S. Congress passed the Smoot-Hawley Tariff Act and sparked a trade war.

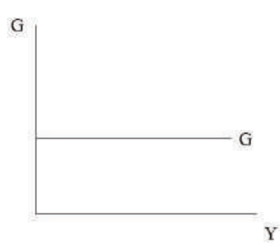
Finally, it is also possible that a change in the foreign exchange value of a nation's currency might alter the trade balance. For example, if the domestic currency depreciates, then the nation's exports will become cheaper for foreigners. The result might be a rise in net exports and a trade surplus. A depreciating currency might then raise output and employment. A potential problem that might arise, however, is retaliatory action taken by foreign central banks. If foreign central banks decide to intervene in the foreign exchange market and deliberately devalue their currencies hoping to acquire a similar competitive trade advantage, then the result might be a net appreciation of the domestic currency. The nation's exports might then become more expensive for foreigners and net exports will fall. This kind of **competitive devaluation** of currencies occurred in the 1930s as well, as different nations struggled to stimulate their domestic economies during the worldwide Great Depression.

Let's now add government spending to the picture by considering the case of an open, mixed economy. A **mixed economy** simply refers to an economy with both a private sector and a public sector. First, we will assume that government spending is exogenously given as follows:

$$G = G_0$$

In other words, government spending is at the same level regardless of the level of real GDP as shown in Figure 13.25.

Figure 13.25: The Government Expenditures Curve

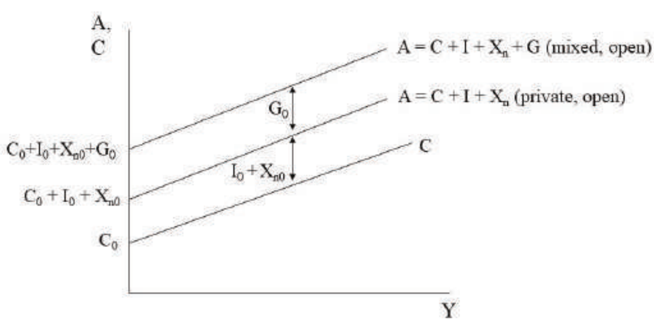


Government spending is assumed to be determined by legislators and a whole host of political factors that neoclassical Keynesian economists do not attempt to explain. Nevertheless, we can now add government spending to the consumption function, the level of planned investment spending, and the level of net exports to obtain the aggregate expenditures function as follows:

$$A = C + I + X_n + G$$
$$A = (C_0 + I_0 + X_{n0} + G_0) + mpc \cdot Y$$

Autonomous expenditure has increased by the amount of the government spending. As Figure 13.26 shows, the addition of government spending increases the vertical intercept of the aggregate expenditures curve by the amount of the government spending.

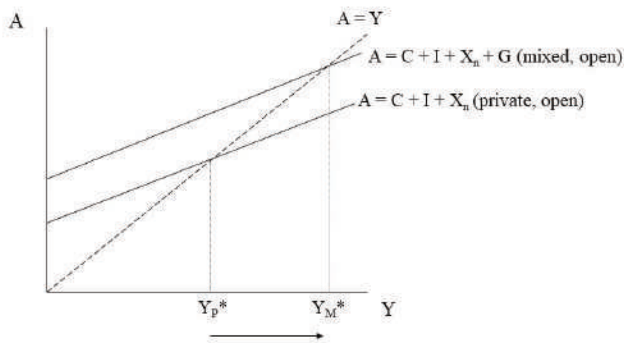
Figure 13.26: Government Expenditures as a Component of Aggregate Expenditures



* This graph assumes positive net exports.

Figure 13.27 shows that the addition of government spending raises the equilibrium real GDP above the level that would exist in a private, open economy.

Figure 13.27: Equilibrium GDP in an Open, Mixed Economy



It should be rather obvious now why Keynes advocated increased government spending during the Great Depression. By increasing government spending, it is possible to increase aggregate output and employment. With a purely private economy stuck at an unemployment equilibrium, government spending can move the economy closer to full employment. Alternatively, reducing government spending during a recession would only worsen the situation by reducing aggregate output and employment in an already weak economy.

We also need to incorporate the other side of the mixed economy, which is the ability of the government to impose taxes. To keep the model relatively simple, we will assume that the government collects a lump sum tax (T) from the households each year. That is, the government does not tax incomes at a particular rate like 20% but rather declares that it will collect a lump sum

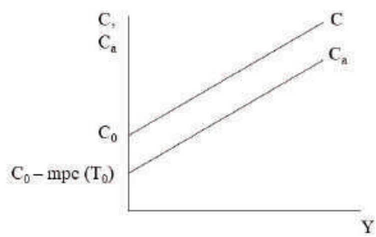
amount of $200 billion from the households regardless of aggregate income. We thus add the following equation to identify this constant amount of taxes collected.
$T = T_0$

Because taxes are collected, it is no longer the case that GDP and disposable income (DI) are equal to one another. To obtain DI, it is now necessary to subtract the lump sum tax from aggregate income (Y). That is, the following equation now holds:
$DI = Y - T$

Since household consumption depends on disposable income, we need to rewrite the consumption function taking into account the lump sum tax. The **after-tax consumption function** is as follows:
$C_a = C_0 + mpc \cdot DI$
$C_a = C_0 + mpc \cdot (Y - T_0)$
$C_a = (C_0 - mpc \cdot T_0) + mpc \cdot Y$

The reader should recall that the **pre-tax consumption function** was the following:
$C = C_0 + mpc \cdot Y$
Therefore, the addition of the lump sum tax causes the consumption function to have a smaller vertical intercept by the amount of the mpc times the lump sum tax, as shown in Figure 13.28.

Figure 13.28: Lump Sum Taxation and After-Tax Consumption



The reason that the lump sum tax lowers consumption at every level of real GDP by this amount is that the households lose the tax amount as part of their income. Because households consume part of their income and save part of their income, when they lose the tax amount, they reduce their consumption by the amount that would have been consumed had they been able to keep this income. That is, they reduce their consumption by the mpc times the amount of the tax.

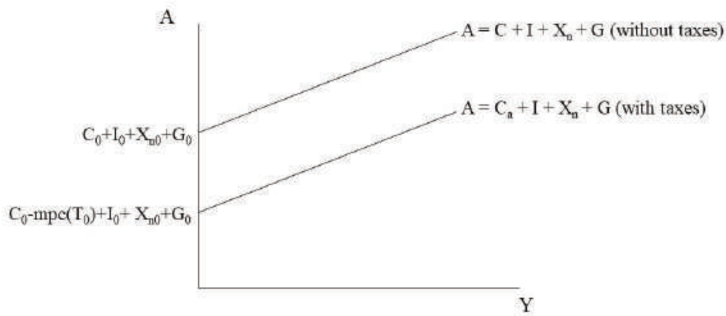
We can now add the after-tax consumption function to the other spending components to obtain the aggregate expenditures function as follows:

$$A = C_a + I + X_n + G$$

$$A = (C_0 - mpc \cdot T_0 + I_0 + X_{n0} + G_0) + mpc \cdot Y$$

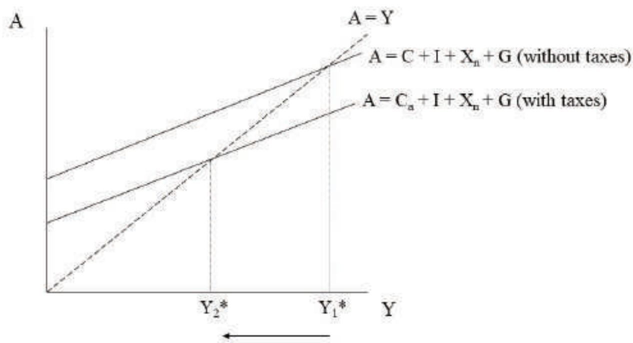
It should be clear that autonomous expenditure has changed yet again. This time it has been reduced by the amount of the mpc times the lump sum tax. The consequence of this change is that the aggregate expenditures curve shifts downward by this amount as shown in Figure 13.29.

Figure 13.29: After-Tax Consumption as a Component of Aggregate Expenditures



Now we can also see the impact that a lump sum tax will have on the equilibrium real GDP. When the lump sum tax is imposed, it shifts the aggregate expenditures curve down, which causes the equilibrium real GDP to fall as shown in Figure 13.30.

Figure 13.30: Equilibrium GDP in an Open, Mixed Economy with Lump Sum Taxation



It is easy to see why a neoclassical Keynesian economist would oppose a tax increase during a recession. Higher taxes discourage consumption which reduces aggregate spending. The drop in spending leads to lower output and employment and thus harms an already weak economy. On the other hand, a tax cut can stimulate consumer spending, which will raise aggregate spending, output, and employment.

The addition of the lump sum tax completes our Keynesian Cross model and allows us to analyze a wide range of possible changes to the aggregate economy. We can also solve algebraically for the equilibrium real GDP if we have enough information. To show how to find this solution, let's assume the following about the economy:

$C_0 = 200$

$I_0 = 100$

$X_{n0} = -100$

$G_0 = 200$

$T_0 = 100$
$mpc = 0.75$

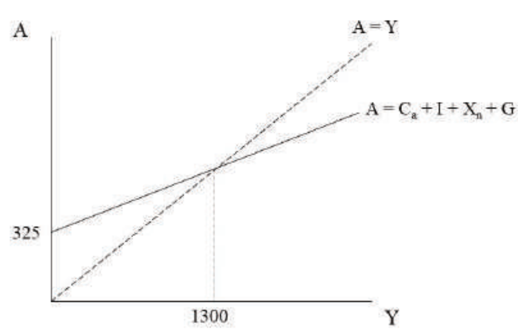
We can also write the complete aggregate expenditures function as follows:
$A = (C_0 - mpc \cdot T_0 + I_0 + X_{n0} + G_0) + mpc \cdot Y$

Plugging in the known information into the aggregate expenditures function yields the following:
$A = (200 - 0.75 \cdot (100) + 100 - 100 + 200) + 0.75 \cdot Y$
$A = 325 + 0.75 \cdot Y$

Now recall the equilibrium condition and solve for Y.
$A = Y$
$325 + 0.75 \cdot Y = Y$
$Y = 1300$

Figure 13.31 provides a graph that corresponds to this solution.

Figure 13.31: The Calculation of Equilibrium Real GDP in the Expanded Keynesian Cross Model



The Lump Sum Tax Multiplier

Just as a change in investment spending leads to a multiplier effect as explained in the last section, it is also possible to identify a multiplier effect stemming from the change in lump sum taxes. The reasoning is similar. When taxes are increased, households lose disposable income. They reduce consumer spending, which causes incomes to fall more. The additional reduction in incomes leads to a great drop in consumer spending, and on and on. As before, with each successive round, consumer spending falls by smaller and smaller amounts because not all the lost income would have been consumed anyway at each step.

It is possible to derive a formula that tells us the exact impact that a change in lump sum taxes has on real GDP, other factors held constant. Let's define the **lump sum tax multiplier** as the ratio of the change in real GDP to the change in lump sum taxes. By taking into account the way in which the households engage in an infinite series of consumer spending rounds, we can prove that the lump sum tax multiplier is negatively related to the marginal propensity to consume as shown below.

$$\frac{\Delta Y}{\Delta T} = \frac{-mpc}{1-mpc}$$

Figure 13A.2 provides the details of the proof for any interested readers. Our main purpose, however, is to understand the intuition behind the formula and to learn how to use it to calculate changes in real GDP. To understand how to use the formula, let's assume that lump sum taxes rise by 100. If the mpc is 0.75, then we can calculate the lump sum tax multiplier by plugging the mpc into the formula as follows:

$$\frac{\Delta Y}{\Delta T} = \frac{-mpc}{1-mpc} = \frac{-0.75}{1-0.75} = \frac{-0.75}{0.25} = -3$$

Because lump sum taxes rise by 100, we can write the equation as follows:

$$\frac{\Delta Y}{100} = -3$$

$$\Delta Y = -300$$

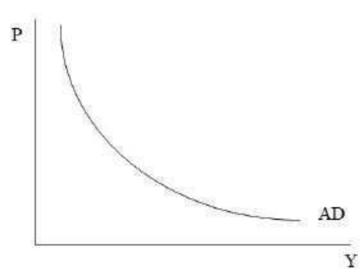
By solving for the change in real GDP, we can demonstrate that real GDP will fall by $300 billion in this case. Alternatively, the multiplier implies that for every $1 of additional taxes, real GDP will fall by $3. Alternatively, a $1 tax cut would lead to a $3 rise in real GDP. Two points are worth mentioning. First, the lump sum tax multiplier is always negative. The reason is that a rise in taxes causes an opposite change in real GDP. This negative relationship exists because higher taxes reduce consumer spending and lower the equilibrium GDP. Second, the lump sum tax multiplier is smaller in absolute value than the investment multiplier.[9] The reader will recall that the investment multiplier was equal to 4 with the same marginal propensity to consume. The reason for the smaller absolute impact of the lump sum tax multiplier is that when the households receive a lump sum tax cut of, say, $100 billion, they will only spend a fraction of it as determined by the mpc. Successive rounds of additional consumer spending then follow. Conversely, when businesses invest an additional $100 billion, the entire $100 billion is spent in the first round, which then leads to successive rounds of additional consumer spending. Because the initial impact of the additional investment spending is larger than the initial impact of the tax cut, the overall impact of an increase in investment spending is significantly larger than the overall impact of a lump sum tax cut.

Aggregate Demand

Up to this point, we have only considered how the levels of aggregate output and employment are determined. In this section, we want to begin building an explanation of the general level of prices. The aggregate demand/ aggregate supply (AD/AS) model was developed for this purpose. The AD/AS model can be understood as an extension of the Keynesian Cross model. We begin by introducing the **aggregate demand curve (**AD) curve and then explain how it relates to the Keynesian Cross model.

The AD curve, shown in Figure 13.32, asserts that an inverse relationship exists between the general price level (P) and the level of real GDP (Y) that is consistent with equilibrium in the market for goods and services.

Figure 13.32: The Aggregate Demand (AD) Curve



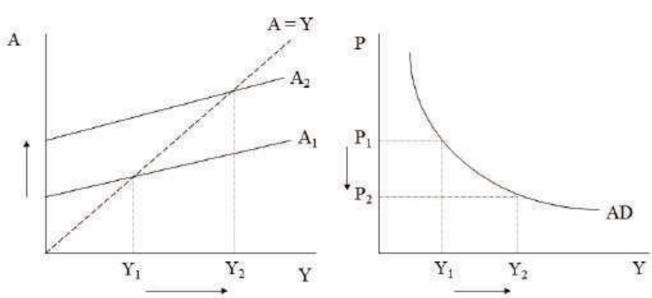
The general level of prices may be thought of as a price index such as the consumer price index or the GDP deflator. Even though the AD curve looks much like a market demand curve, it is actually quite different. It turns out that the law of demand does not apply in this context.[10] When we discussed the market demand curve in Chapter 3, it was argued that the market demand curve slopes downward for two main reasons.

First, when the price of an individual commodity falls, consumers substitute away from relatively more expensive commodities whose prices have not changed. This effect, which causes a movement along the demand curve, was referred to as **the substitution effect**. The downward slope of the AD curve cannot be explained in a similar fashion. When the general price level falls, for example, all commodity prices in the economy are falling and so it does not make sense to talk about substitution away from relatively more expensive domestically produced commodities. It is true that deflation does not necessarily mean that all prices are falling at the same rate. Nevertheless, a drop in a price index does not allow us to detect variation in the reduction of prices across commodities and so this explanation will not suffice as an explanation of the downward sloping AD curve.

Second, when the price of an individual commodity falls, consumers experience a rise in their real incomes. That is, the purchasing power of their nominal incomes increases. Feeling richer, they increased their quantity demanded of the commodity whose price fell as well as the quantities demanded of all other commodities. This effect, which also contributes to the movement along the demand curve, was referred to as **the income effect**. The downward slope of the AD curve cannot be explained in a similar fashion. When a period of generalized deflation occurs, for example, input prices fall in addition to product prices. The result is that factor incomes decline. With nominal incomes declining along with commodity prices, real incomes are likely to remain the same on average. Therefore, we should not expect an income effect at the aggregate level.

Because the law of demand cannot explain the downward slope of the AD curve, we require a different explanation for its downward slope. To understand its shape, we will first explain the relationship of the AD curve to the Keynesian Cross model. Suppose that the price level falls. We will claim, for reasons not yet explained, that this drop in the price level causes aggregate expenditure to rise as shown in Figure 13.33.

Figure 13.33: The Derivation of the Downward-Sloping Aggregate Demand Curve



As shown in Figure 13.33, the aggregate expenditures curve shifts upward and raises the level of equilibrium real GDP. The consequence is a negative relationship between the general price level and the level of real GDP. The AD curve thus slopes downward. An explanation must be provided, of course, for the negative relationship between the price level and aggregate expenditure. Neoclassical Keynesian economists provide three main explanations for this negative relationship.[11]

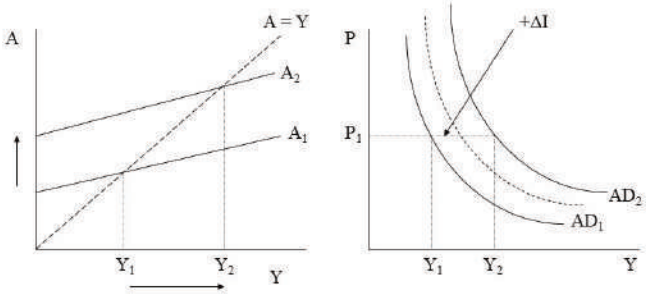
The first explanation for the negative relationship between aggregate expenditures and the general price level is referred to as the **wealth effect** or the Pigou effect after the classical economist, A.C. Pigou. According to this line of thinking, when the price level falls, even though households do not experience a rise in their real incomes, they do experience a rise in their real wealth. Because other factors are held constant, including nominal wealth (e.g., home prices, stock prices), households experience a rise in the purchasing power of their wealth. As a result, they increase their consumption expenditures, which stimulates aggregate expenditure and raises the equilibrium real GDP. The result is a downward sloping AD curve.

The second explanation for the negative relationship between aggregate expenditures and the general price level is referred to as the **international substitution effect**. According to this line of thinking, when the price level falls, even though no substitution away from relatively more expensive domestically produced commodities occurs, substitution away from relatively more expensive foreign commodities does occur. That is, the drop in the general price level only refers to domestically produced commodities with everything else remaining constant, including prices of foreign commodities. As a result, imports decline and net exports rise. Exports also rise because foreign buyers now substitute towards relatively cheaper commodities in this nation. The aggregate expenditures curve thus shifts upward and the equilibrium real GDP rises. The result is a downward sloping AD curve.

A final explanation for the negative relationship between aggregate expenditures and the general price level is referred to as the **interest-rate effect**. It is also sometimes referred to as the Keynes effect because Keynes was the first to identify it. According to this effect, when the price level falls, people decide to hold less money because they need less money to engage in transactions. As a result, they lend their excess money holdings, which pushes down the rate of interest. The fall in the rate of interest stimulates investment spending and raises aggregate expenditure. As a result, equilibrium real GDP rises. The result is a downward sloping AD curve.

We are now able to discuss the factors that tend to shift the AD curve. Suppose that for a given price level of P_1, the economy is at an equilibrium real GDP of Y_1 in the Keynesian Cross model as shown in Figure 13.34.

Figure 13.34: Determinants of Aggregate Demand: An Example



Now suppose that planned investment spending rises. The aggregate expenditures curve shifts upward, which raises the equilibrium real GDP to Y2. Because the price level has not changed, the equilibrium real GDP will be higher at the same price level in the graph of the AD curve. This change implies a movement off of the AD curve and to the right. Because such movements to the right would occur at any given price level when the level of investment rises, it should be clear that the entire AD curve shifts rightward when investment spending increases. The reader might also note that the AD curve shifts rightward by more than the amount of the increase in investment spending due to the multiplier effect, which is consistent with Figure 13.20. That is, the change in equilibrium output at the current price level more than exceeds the change in investment spending.

Although this example concentrates on a shift of the AD curve due to a change in investment spending, a change in any component of aggregate expenditure will have a similar impact on the position of the AD curve. In general, a rise in consumer spending, investment spending, government spending, or net export spending will shift the AD curve rightward. Similarly, a reduction in consumer spending, investment spending, government spending, or net export spending will shift the AD curve leftward.

To be more specific, consider factors that might influence consumer spending. A rise (fall) in nominal wealth will stimulate (depress) consumer spending and shift the AD curve rightward (leftward). The reader should notice that this effect is not the same as the wealth effect that produced a downward sloping AD curve. The reason is that in this scenario, it is a change in nominal wealth that causes a change in real wealth, rather than a change in the general price level that causes a change in real wealth. Another factor that might alter consumer spending is a change in taxes on household income. If taxes are reduced, then this change will stimulate consumer spending and raise aggregate expenditure. The higher equilibrium real GDP will show up as a shift of the AD curve to the right. A tax increase would have the opposite effects.

Changes in investment spending are likely to have different causes. One major factor influencing investment spending is the rate of interest. If the rate of interest falls, then businesses will borrow more because they are more likely to profit from new investment projects. The rise in investment spending will raise aggregate expenditure and equilibrium real GDP. The result will be a rightward shift of the AD curve. A rise in the interest rate would have the opposite impact and lead to a leftward shift of the AD curve. Other factors that might affect the level of investment include changes in expected profitability. The expected profits from new investment might change to due to changes in the state of the economy, changes in production technology, or changes in business taxes. If business expectations improve, new technologies are developed, or business taxes are cut, then expected profits rise, investment rises, aggregate expenditure rises, and the AD curves shifts rightward. A reduction in expected profits due to the opposite conditions would shift AD to the left.

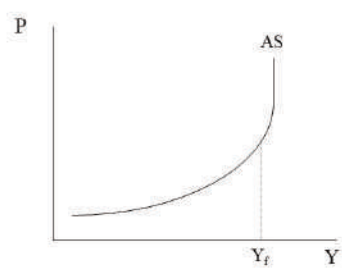
A change in government spending has a direct effect on the position of the AD curve as well. If government spending rises, then aggregate expenditure rises. The equilibrium real GDP rises, and the AD curve shifts rightward. If government spending falls, then the opposite effects occur, and the AD curve shifts leftward.

A change in net export spending will also influence the position of the AD curve. If trading partners experience economic expansions and incomes are rising, then net exports will rise, raising aggregate expenditures and equilibrium real GDP. As a result, the AD curve will shift rightward. If trading partners experience recessions, then the effects are the opposite and the AD curve shifts leftward. Changes in tariff policy and changes in the foreign exchange value of the domestic currency may also shift the AD curve, although the effects are uncertain due to the possibility of retaliatory tariffs or competitive currency devaluation. Without retaliation, the imposition of tariffs or the devaluation of the currency will discourage imports, raise net exports, raise aggregate spending, and raise equilibrium real GDP. The consequence will be a rightward shift of the AD curve. A reduction in tariff rates or an appreciation of the domestic currency would have the opposite effect if trading partners do not alter their policies, and the AD curve would shift leftward.

Aggregate Supply

Our discussion of aggregate demand has suggested that aggregate spending is the primary determinant of the amount of output produced in the economy. It seems to ignore one other factor that arguably plays a major role in the determination of output: production cost. To capture the role of cost of production in the determination of aggregate output, we turn to the aggregate supply side of the economy. Figure 13.35 shows a graph of the **aggregate supply curve (**AS) curve.

Figure 13.35: The Aggregate Supply (AS) Curve and Rising Per-Unit Production Costs

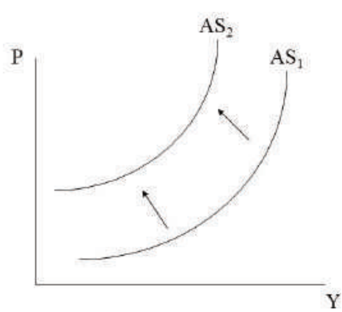


The graph suggests that a positive relationship exists between the general price level and the level of real GDP that businesses are willing and able to produce at each price level. The aggregate supply (AS) curve looks much like a market supply curve, and the explanation of its shape is similar. That is, as the price level rises, per unit profit rises and so firms expand production, but the increase in production drives up unit costs, which brings the expansion to a halt unless the price level rises further. As production rises, the per unit cost of real output rises due to diminishing returns to labor. The price level must rise to cover the higher per unit cost.

Referring to Figure 13.35, when the economy is operating below the full employment GDP (Y_f), increases in real GDP do not put much upward pressure on input prices or unit costs due to the great deal of excess capacity in the economy. As a result, the general price level does not tend to rise much. As the economy approaches the full employment level of GDP, however, efficient resources become more and more difficult to acquire.[12] As a result, less efficient resources must be hired and unit production costs begin to rise. The general price level must, therefore, rise to compensate for the higher unit production costs. That is, businesses raise prices as their production costs per unit increase. A related reason for the rise in per unit production costs as real GDP rises has to do with diminishing returns to labor. With the aggregate stocks of capital and land being relatively fixed during this relatively short time period, the increase in employment raises production but at a decreasing rate. Therefore, it is necessary to hire increasing numbers of workers to raise the production of real GDP by one unit. Hence, unit production costs also rise for this reason and contribute to the upward slope of the AS curve.

The next question that arises deals with the factors that shift the AS curve. Basically, a change in any variable that affects per unit production cost, other than a change in the level of real GDP, will shift the AS curve. A rise in per unit production cost will shift the AS curve to the left because at each level of real GDP, the prices that firms require will need to be higher. A fall in per unit production cost will shift the AS curve to the right because at each level of real GDP, the prices that firms require will be lower. For example, a change in the nominal wage rate will shift the AS curve. It is assumed that when the price level changes, all other variables are held constant, including input prices like wages. If the nominal wage rate rises, then at any given level of real GDP, per unit costs will be higher. The result is a leftward shift of the AS curve as shown in Figure 13.36.

Figure 13.36: Determinants of Aggregate Supply: An Example



Alternatively, a reduction in the nominal wage would shift the AS curve to the right because per unit cost would be lower at every level of real GDP.

A variety of other factors will also shift the AS curve, but each factor works by influencing the per unit production cost of firms. For example, if the nominal prices of land and capital rise, then the AS curve will shift leftward because unit costs are higher. If other input prices fall, then the AS curve will shift rightward.[13] Factor supplies will also affect the position of the AS curve. If factor supplies (i.e., the supplies of land, labor, and capital) rise, then input prices will fall and the AS curve will shift to the right. If the factor supplies fall, then input prices will rise and the AS curve will shift to the left.[14]

Other factors include the prices of imported inputs.[15] For example, if the price of imported oil rises, then unit production costs will rise and shift AS to the left. If import prices fall, then the AS curve will shift to the right. Changes in the foreign exchange value of the domestic currency will also affect unit cost by making imported inputs more or less expensive. For example, if the domestic currency appreciates, then imported inputs become cheaper for domestic producers. Their unit costs fall, and the AS curve shifts to the right. If the domestic currency depreciates, then imported inputs become more expensive for domestic producers and unit costs rise. The AS curve would then shift to the left.

Changes in labor productivity are also likely to affect unit cost and the position of the AS curve.[16] Labor productivity is measured in terms of output per worker (Q/L) where Q is the number of units of output and L is the number of employed workers. Labor cost per unit may be measured in terms total wages per unit produced (wL/Q) where w is the wage rate. It should be clear that a rise in productivity (Q/L) will cause a drop in labor cost per unit (wL/Q). Hence, a rise in labor productivity will shift the AS curve to the right. If productivity falls, then unit labor cost will rise and shift the AS curve to the left.

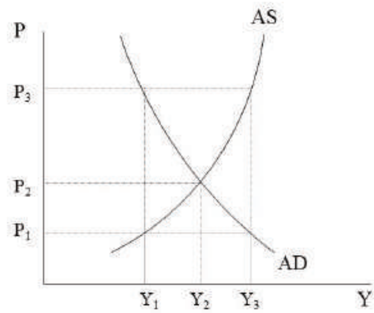
The degree of monopoly power in input markets may also influence unit cost.[17] Because monopoly markets produce higher prices than competitive markets, if monopoly power grows in a major input market, then unit costs rise, and the AS curve shifts to the left. If government antitrust action breaks up a monopoly in an input market, on the other hand, then input prices will fall and unit costs will fall as well. The AS curve would then shift to the right.

Finally, changes in business taxes and tax credits may also influence per unit production cost.[18] If business taxes are cut, then unit cost will fall, and the AS curve will shift to the right. If business taxes are increased, then unit labor cost will rise and the AS curve will shift to the left. Alternatively, if the government increases its subsidies or tax credits for business, then in effect unit cost will fall, and the AS curve will shift to the right. If the government cuts its subsidies or tax credits to business then in effect unit cost will rise, and the AS curve will shift to the left.

Macroeconomic Equilibrium and Historical Applications of the Model

Now that the aggregate demand and aggregate supply sides of the model have been developed, we can combine them to explain how **macroeconomic equilibrium** occurs. Macroeconomic equilibrium occurs when the general price level and the level of real GDP reach levels from which there is no inherent tendency to change. Figure 13.37 depicts an economy that has reached a macroeconomic equilibrium state where P_2 and Y_2 are the equilibrium general price level and the equilibrium real GDP, respectively.

Figure 13.37: Macroeconomic Equilibrium

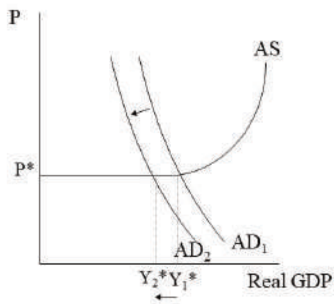


We want to ask how the economy reaches the equilibrium outcome. It may be tempting to fall back on the explanation that we offered when we discussed the movement to equilibrium in an individual market as in Chapter 3. In this model, however, we need to refer to the factors that cause movements along the AD and AS curves. For example, suppose that the price level is at P_3 in Figure 13.37. In this case, aggregate spending is relatively low compared with what businesses wish to produce at this price level. If businesses are producing Y_3 and aggregate spending leads to a lower equilibrium GDP in the Keynesian Cross model, then firms will experience a rise in inventories (unplanned investment). They will then cut production. As they cut production, a movement down along the AS curve occurs. The release of relatively inefficient resources causes unit production cost to fall and the price level begins to decline. As the price level falls, three effects on the aggregate demand side occur that stimulate aggregate expenditure. When P falls, households experience a rise in real wealth and consume more. Also when P falls, foreigners begin to buy more of this nation's exports and domestic buyers buy fewer imports. These factors increase net export spending. Finally, when P falls, the amount of money people wish to hold falls, which leads to more lending, lower interest rates, and higher investment spending. All three factors cause a movement down along the AD curve. Eventually, the economy will arrive at the macroeconomic equilibrium outcome.

Alternatively, suppose that the economy begins with a price level of P_1. In this case, businesses do not want to produce much real output given the low price level that just covers their low unit costs. On the other hand, aggregate spending is rather high, leading to a high equilibrium GDP in the Keynesian Cross model. Aggregate output will begin to rise as inventories are depleted as a result of the high aggregate spending. As output begins to rise, the effects of diminishing returns to labor and the employment of less efficient resources causes unit costs to creep upwards. Businesses will raise prices to keep up with rising unit costs. As prices rise, three effects on the aggregate demand side begin to take effect. As P rises, households experience a reduction in their real wealth, which leads to lower consumer spending. As P rises, foreigners substitute away from the nation's exports and domestic buyers substitute towards imports. Both effects reduce net export spending. Finally, as P rises, people wish to hold more money for transactions purposes and so they lend less. The result is a rise in the rate of interest, which discourages investment spending. All three effects cause a movement up along the AD curve. Eventually, the economy will arrive at the macroeconomic equilibrium outcome.

The AD/AS model can be easily applied to specific periods in U.S. macroeconomic history to obtain a sense of how and why the price level and level of real GDP changed. For example, prior to World War II, the U.S. economy frequently experienced periods of deflation during recessions. The recessions that have occurred since World War II have generally not been characterized by deflation. The AD/AS model can help us to understand why deflation has become less common in the U.S. economy. Figure 13.38 shows a horizontal section of the AS curve, which implies that when AD declines during a recession, the price level does not fall.[19]

Figure 13.38: A Horizontal Aggregate Supply Curve

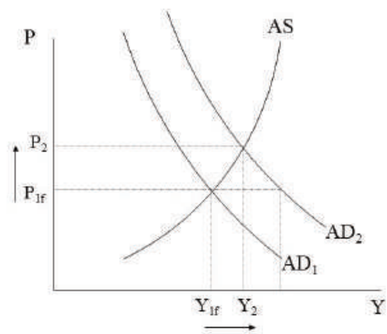


During these types of recessions, only real GDP declines. To understand why a horizontal aggregate supply curve might exist, we need to consider how conditions changed in the 1930s and 1940s. In particular, it became more difficult to reduce wages during a recession for a number of reasons, including the passage of a federal minimum wage law in 1938 and the growing power of industrial unions in major industries that negotiated high wage rates. With unit labor costs becoming more rigid, businesses could not reduce them during a recession and so they hesitated to cut prices. Other factors contributing to **sticky prices** might have been business concerns about **menu costs** (costs associated with price cuts) and efforts to avoid price wars with oligopolistic competitors.[20] The tendency for prices to be downwardly inflexible means that prices tend to increase, but once they rise, they tend not to fall again.

Let's consider what occurred in the U.S. during the

1960s. This decade was marked by rapid economic growth but also a rising price level. Factors that contributed to these changes were rising government spending related to America's involvement in the Vietnam War and President Johnson's poverty reduction programs. The rise in government spending led to a rightward shift of the AD curve as shown in Figure 13.39.

Figure 13.39: Case 1 - Demand-Pull Inflation

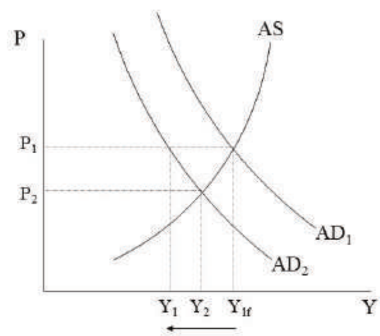


The rise in aggregate demand increased the general price level and the level of real GDP. Because the economy was near the full employment GDP (Y_{1f}), the rise in aggregate demand pushed the unemployment rate below the natural rate of unemployment and had a strong inflationary impact. When inflation is the result of a rise in aggregate demand, economists generally refer to it as a case of **demand-pull inflation**.[21]

Next let's consider what occurred in the U.S. during the

1930s. This decade was characterized by a severe reduction in real GDP as well as a falling price level or deflation. The main factor contributing to these changes was a drop in aggregate demand as shown in Figure 13.40.[22]

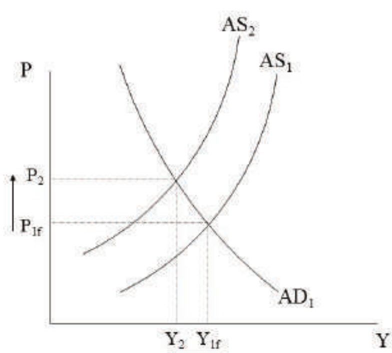
Figure 13.40: Case 2 – The Great Depression



After the stock market crash of 1929, investment spending and consumer spending fell significantly. The failures of many American banks contributed significantly to the reduction in investment spending as well. As real output fell and unemployment rose, unit costs declined and the price level fell. The price level and level of real output eventually reached P_2 and Y_2, respectively. If the economy had been subject to downwardly inflexible prices, then the price level would have remained at P_1 and the real output would have fallen to Y_1. Interestingly, such price stickiness would have made the collapse of output and employment *worse* during the Great Depression. The falling price level actually took some of the pressure off of real output as aggregate demand fell. The reason that recessions since World War II have been less severe is not the result of sticky prices. Instead, it is because reductions in aggregate demand have been smaller as governments and central banks have become more aggressive about preventing economic collapses.

During the 1970s, the U.S. economy experienced the unpleasant coincidence of a rising price level and falling real output. Figure 13.41 shows that a leftward shift of the AS curve was responsible for this inflationary recession.

Figure 13.41: Case 3 - Cost-Push Inflation

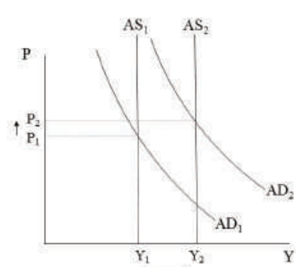


When inflation occurs at the same time as falling real GDP, the situation is referred to as **stagflation**. It is the worst of both worlds (i.e., inflation and recession). The reason that the AS curve shifted leftwards in the 1970s had to do with a sudden rise in unit production cost.

Unit costs rose because of the two oil price shocks that increased the price of imported oil in the United States. In 1973, the Arab-Israeli War broke out which interrupted the flow of oil to the West. Then in 1979 the revolution in Iran once again led to a disruption of the flow of oil westward. In both cases, tighter global supplies caused the price of oil to skyrocket, raising production costs and leading to stagflation. Because the inflation in this case was the result of an aggregate supply shift, it is often referred to as **cost-push inflation** to distinguish it from the demand-pull inflation of the 1960s.[23]

Another interesting case occurred during the 1990s. The 1990s represents the longest economic expansion in U.S. economic history. Unlike the 1970s, this decade captured the best of both worlds with both rapid growth of real GDP and a relatively stable price level. Figure 13.42 provides insight into the factors that contributed to this situation. The figure assumes a vertical aggregate supply curve for reasons that are explained in the next section. 24

Figure 13.42: Case 4 - Full Employment with Stable Prices



As Figure 13.42 shows, aggregate demand rose during this decade. A major reason for the rise in AD was the information technology (IT) boom that raised the expected profitability of many startup companies. These changes led to a surge of investment in the IT sector. With IT stocks rising in value, households also experienced a rise in nominal and real wealth, which led to a rise in consumer spending. While these changes would normally cause demand-pull inflation, like the kind observed in the U.S. in the 1960s, in this case a rise in aggregate supply occurred simultaneously. The reason was that the technological changes in the IT sector also raised labor productivity, which reduced unit labor costs and shifted the AS curve to the right. As Figure 13.42 indicates, both the rise in AD and the rise in AS contributed to a large increase in real GDP. At the same time, only a small increase in the price level occurred because the rise in AS put downward pressure on prices even as the rise in AD tended to push prices up. The economy of this decade was dubbed "The New Economy," which included the claim that the business cycle had been conquered and that uninterrupted and stable economic growth was to be expected forevermore. The experience of the next decade would serve as a reminder that the business cycle has been far from conquered.[25]

In the year 2000, it became clear that the good times had ended. A stock market bubble had been forming in the IT sector for years. **Asset price bubbles** form when asset prices rise significantly above levels that are consistent with the real underlying values of the assets. In the case of the IT sector, "irrational exuberance" had led to the bidding up of IT stocks far beyond what would be justified by the profitability of the associated companies. When investors finally discovered to what extent the stocks were overvalued, they dumped them, and the prices collapsed. This disruption in the stock market meant that a great deal of paper wealth was destroyed in a short time. The result was a collapse of investment and consumer spending. Aggregate demand thus fell and gave way to the recession of 2001, like what is depicted in Figure 13.38.

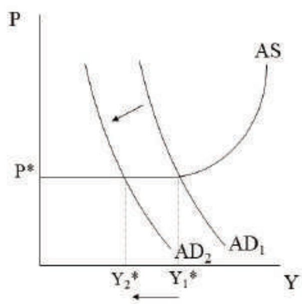
The 2001 recession was mild, however, and the economy soon recovered. The nation's central bank, the U.S. Federal Reserve, which controls the nation's money supply, took the lead in responding to the crisis. It increased the money supply which pushed interest rates down to very low levels. As interest rates fell, investment spending rose, but this time investments flowed into the housing sector rather than the IT sector. Over the course of the next few years, a boom in home construction occurred. Low mortgage interest rates encouraged the purchase of new homes, which led to soaring home prices. As before, the bidding upward of these prices above their underlying real values implied the formation of an asset price bubble.

A factor that contributed greatly to the formation of this asset price bubble was the way large financial institutions encouraged the growth of the market for various financial assets, including so-called **mortgage-backed securities** (MBSs). When commercial banks grant mortgage loans to home buyers, they typically do so for a period of 15 or even 30 years. In the past, a commercial bank would carefully scrutinize the credit worthiness of the home borrower before granting the loan to ensure that the bank would receive mortgage payments each month until the loan was repaid with interest. The growth of the market for MBSs meant that a commercial bank could sell this loan to a large financial institution, like Goldman Sachs, which would then bundle together dozens or even hundreds of such mortgage loans that originated in many different places, creating a single financial asset (i.e., a mortgage-backed security). Goldman Sachs would then sell the MBS to a large institutional investor like a hedge fund. The asset might be sold many times in the organized market for these securities that arose. The owner of the MBS would receive interest payments from many different homeowners due to its ownership of the asset.

The problem that this situation created was that so much money was being made by packaging and selling these specialized assets that less attention was being paid to the credit worthiness of the borrowers. When the loan originator (i.e., the bank that initially grants the loan) is not the same institution that will suffer if the borrower defaults on the loan, it is much less likely that the loan originator will take the proper care in evaluating the credit worthiness of the borrower. Furthermore, credit rating agencies, like Moody's and Standard and Poor's, which were supposed to signal to investors how much risk they were assuming by purchasing these securities encountered a conflict of interest. They collected more fees by rating more of these securities. Positive ratings were likely to encourage the growth of the market and keep the market for these securities active. As a result, these agencies tended to be far too optimistic in their valuations and tended to underestimate the degree of risk associated with these securities.

The consequence of all these factors was that home buyers eventually began to default on loans in large numbers. These defaults triggered a collapse in the prices of mortgage-backed securities. Financial institutions that held these assets on their balance sheets watched as the losses mounted. Driven by fear of additional losses, they contracted their lending. The contraction of lending reduced investment spending. Many businesses were unable to obtain the loans necessary simply to pay employees, and business failures began to increase. The result was a huge collapse of aggregate demand as shown in Figure 13.43.

Figure 13.43: Case 5 – The Great Recession



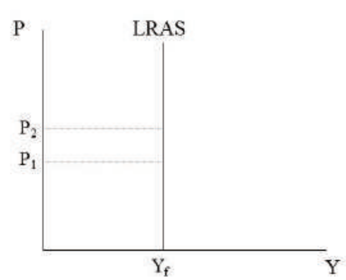
This recession was so extreme that it has been dubbed the **Great Recession**. It was the worst reduction in real GDP that has occurred since the Great Depression of the 1930s. It should be noted, however, that the Great Depression was far worse with a drop in real GDP and an unemployment rate in excess of 25%. By contrast, during the Great Recession, real GDP fell by just over 4% and the unemployment rate was around 10% at its peak. The reason that the Great Recession was not worse than it was stemmed from the massive government and central bank responses to the crisis. The U.S. Federal Reserve acted as a lender of last resort and offered emergency loans to large financial institutions. The federal government also passed emergency measures, including a $700 billion Troubled Asset Relief Plan (TARP) in late 2008 that bought up hundreds of billions of dollars of so-called toxic assets from banks and other financial institutions while also purchasing equity stakes in the same financial institutions. The federal government also implemented a fiscal stimulus package in early 2009 that included $787 billion of government spending increases and tax cuts. This stimulus package was consistent with the Keynesian prescription for boosting output and employment during recessions. Although these measures had some impact, economic growth remained sluggish for several years and unemployment remained stubbornly high.

The Neoclassical Synthesis Model and the Post-Keynesian Critique

The model that we have investigated up to this point is consistent with the neoclassical interpretation of Keynes's theory. Keynes's *General Theory* is a difficult book and is subject to numerous interpretations. After World War II, Paul Samuelson identified something called the **Neoclassical Synthesis model**. This model aims to capture what is most valuable in Keynes's theory while also retaining much of the classical or early neoclassical model that Keynes rejected as incomplete. Because it retains so much of the early neoclassical perspective, many Post-Keynesian economists, who adhere to a very different interpretation of Keynes's theory, reject the neoclassical synthesis model. We will consider more of the Post-Keynesian perspective in the next chapter. In this section, however, we will briefly consider how neoclassical Keynesian economists aim to create a merger or synthesis of neoclassical and Keynesian ideas.

The proposed merger essentially rests on a distinction between the short run and the long run. As the argument goes, Keynes's theory best applies to short run changes in aggregate output and employment when nominal wages are sticky. In the long run, however, when nominal wages and other input prices are flexible, it is the classical theory that best applies. To understand this argument, we need to develop a **long run aggregate supply curve** (LRAS) curve to distinguish it from the upward sloping **short run aggregate supply curve** (SRAS) curve that we have been using throughout this chapter. The LRAS curve is shown in Figure 13.44.

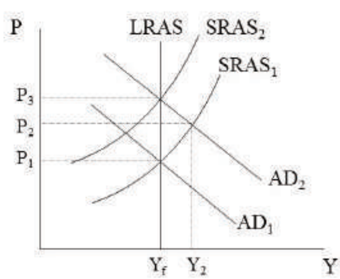
Figure 13.44: The Long Run Aggregate Supply (LRAS) Curve



The LRAS curve is perfectly vertical. The reason is that in the long run with perfect price flexibility, all resources will be fully employed. Any unemployment in the markets for land, labor, or capital, will disappear as prices fall and eliminate the unemployment. With each factor of production fully employed, the economy will produce the full employment level of real GDP. Changes in the price level will have no impact on the level of real GDP. The only factors that can shift the LRAS curve are changes in the endowment of society's resources (i.e., land, labor, and capital) or changes in the available production technology. Otherwise, the LRAS curve will remain where it is.

Now let's begin at a point of long run macroeconomic equilibrium as shown in Figure 13.45 where the AD, SRAS, and LRAS curves all intersect to produce an equilibrium price level of P_1 and an equilibrium real output of Y_f.

Figure 13.45: Demand-Pull Inflation in the Neoclassical Synthesis Model



Now suppose that the central bank increases the money supply, which pushes the interest rate down and stimulates investment spending. The rise in investment spending increases aggregate demand, shifting the AD curve to the right. According to Keynesian theory, output rises above the full employment level to Y_2 and the unemployment rate falls below the natural rate of unemployment. At the same time, the price level rises to P_2 due to the rise in per unit cost. The Keynesian short-run explanation would stop at this point, but according to the neoclassical synthesis model, in the long run factor prices will begin to rise. As nominal wages and other factor prices rise pushing up unit cost, the SRAS curve shifts to the left. The new long run equilibrium then occurs at the intersection of AD, the new SRAS curve, and the LRAS curve. Output returns to the long run level, and the price level is permanently higher at P_3. Even though Keynes's theory provides an explanation for the short run fluctuations, the classical theory provides the long run explanation.

Of course, Post-Keynesian economists are highly critical of the neoclassical synthesis model. The model assumes that if we wait long enough, then wages will adjust to bring about full-employment. It was precisely this kind of thinking that Keynes rejected in his *General Theory*. Keynes's famous remark in response to such thinking was that, "In the long run, we are all dead." Furthermore, this particular example suggests that money is not neutral in the short run, but it is neutral in the long run. Those who defend the **neutrality of money** argue that changes in the money supply cannot cause changes in real variables, like output and employment. Post-Keynesian economists have long argued that Keynes's theory implies the non-neutrality of money in both the short run and the long run. More will be said about the Post-Keynesian perspective in later chapters.

Following the Economic News[26]

Economic news articles frequently refer to changes in the overall economy. With quarterly estimates of real GDP, economists focus on the growth rate of real GDP from one quarter to the next. An article by Eric Morath in *The Wall Street Journal* from July 28, 2016 titled, "5 Things to Watch in the Second-Quarter GDP Report," states that real GDP is estimated to have increased at an annualized rate of 2.6% in the second quarter of 2016. This rate of economic growth exceeded the first quarter growth rate of 1.1%. According to the article, several factors prevented the rate of economic growth from being higher. The relatively strong U.S. dollar and weak overseas economies reduced U.S. exports. As we learned in this chapter, these factors shift the aggregate expenditures curve downward, reducing the equilibrium level of real GDP in the Keynesian Cross model. They also shift the AD curve to the left, reducing output and reducing inflationary pressures (although not causing deflation in the presence of sticky prices). The stronger dollar might also shift the AS curve to the left because it raises the cost of imported inputs and raises the per unit production costs of firms. The article also refers to a rise in imports, which reduces net exports and would further dampen aggregate demand. Trade may have thus served as "a drag on overall growth." The rise in the rate of economic growth is attributed to "relatively healthy consumer spending" between April and June of 2016. As we learned, higher consumer spending raises aggregate spending and aggregate demand, both of which lead to a higher equilibrium level of real GDP even though they leave the price level relatively unchanged in the presence of inflexible prices. Finally, it is argued that stronger growth may have been supported as businesses added to their inventories in the spring. Here it is important to remember that inventory investment may be planned. A rise in planned inventory investment would then shift the aggregate expenditures curve upwards and the AD curve to the right. It is only a rise in unplanned inventory investment that leads firms to cut production.

Summary of Key Points

1. Say's Law of Markets states that new production will always generate enough income for its purchase and so general gluts of overproduction are not possible in market capitalist economies.
2. According to J.M. Keynes, Say's Law of Markets does not hold in market capitalist economies because not all savings will be invested.
3. The consumption function and the saving function show that consumer spending and saving are positively related to the disposable income of households.
4. The Keynesian Cross model explains how the economy arrives at an equilibrium level of real GDP as businesses change production in response to a buildup or depletion of inventories.
5. In the Keynesian Cross model, the equilibrium condition states that aggregate planned expenditure is equal to real GDP.
6. The paradox of thrift asserts that even if all households increase their saving, aggregate saving will ultimately not change.
7. The Keynesian multiplier effect refers to the

tendency for real GDP to rise by a multiple of a rise in investment spending.

8. In the Keynesian Cross model, trade surpluses raise equilibrium GDP, trade deficits reduce equilibrium GDP, and balanced trade leaves equilibrium GDP unchanged.
9. In the Keynesian Cross model, government spending increases or tax cuts raise equilibrium real GDP, whereas government spending reductions or tax increases reduce equilibrium real GDP.
10. The lump sum tax multiplier is negative and has a smaller absolute impact on real GDP than the investment multiplier.
11. The AD curves slopes downward due to the wealth effect, the international substitution effect, and the interest-rate effect.
12. The AD curve shifts due to changes in consumer spending, investment spending, government spending, and net exports.
13. The AS curve slopes upward due to rising unit costs as less efficient resources are employed and diminishing returns to labor occurs.
14. The AS curve shifts due to changes in all other factors that affect unit costs, such as changes in input prices, import prices, the exchange rate, business taxes and credits, monopoly power, labor productivity, and input supplies.
15. Macroeconomic equilibrium occurs at the intersection of the AD and AS curves and determines the equilibrium real GDP and price level.
16. The neoclassical synthesis model represents a

merger of neoclassical economics and Keynesian economics.

List of Key Terms

Say's Law of Markets

Aggregate production function

Consumption function

Autonomous consumption

Marginal propensity to consume (mpc)

Saving function

Autonomous saving

Marginal propensity to save (mps)

Reference line

Keynesian Cross model

Aggregate expenditures model

Induced spending

Planned investment

Actual investment

Paradox of thrift

Induced saving

Fallacy of composition

Multiplier effect

Animal spirits

Trade surplus

Trade deficit

Balanced trade

Tariff

Competitive devaluation

Mixed economy

After-tax consumption function

Pre-tax consumption function

Lump sum tax multiplier

Aggregate demand (AD) curve

Substitution effect

Income effect

Wealth effect

International substitution effect

Interest-rate effect

Aggregate supply (AS) curve

Macroeconomic equilibrium

Sticky prices

Menu costs

Demand-pull inflation

Stagflation

Cost-push inflation

Asset price bubbles

Mortgage-backed securities (MBSs)

Great Recession

Neoclassical Synthesis model

Long run aggregate supply (LRAS) curve

Short run aggregate supply (SRAS) curve

Neutrality of money

Problems for Review

1.Suppose the consumption function is C = 250+0.8DI. What is the saving function?

2. Suppose that investment spending falls by $300 billion. If the mpc is 0.6, then what is the change in real GDP, according to the Keynesian multiplier effect?

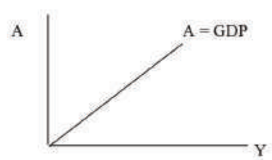
3. Suppose you are given the following information about the economy:

- Investment spending is 400
- Government spending is 600
- A trade deficit of 250 exists

- Autonomous consumption is 300
- The mpc is 0.70
- A lump sum tax of 275 is imposed

Given this information, write the aggregate expenditures function. Then calculate the equilibrium level of real GDP and place your answer on a graph like the one below.

Problem #3



4. Suppose that a lump sum tax cut of $125 is imposed on households. If the mpc is 0.82, then what will be the overall change in real GDP that results?

5. Suppose that the economy begins in macroeconomic equilibrium as depicted in the AD/AS model. Suppose that a stock market collapse occurs that reduces household wealth at the same time that monopoly power increases in key input markets. What will happen to the equilibrium levels of prices and real GDP? Represent your answer graphically.

6. Suppose that the economy begins in macroeconomic equilibrium as depicted in the AD/AS model. Suppose

that the domestic currency depreciates. What will happen to the equilibrium levels of prices and real GDP? Consider the impact on both AS and AD in your answer. Represent your answer graphically.

7. Suppose that the economy begins in macroeconomic equilibrium as depicted in the AD/AS model. Suppose that labor productivity rises at the same time that taxes on households are reduced. What will happen to the equilibrium levels of prices and real GDP? Represent your answer graphically.

Figure 13A.1: The Investment Multiplier and the Marginal Propensities to Consume and Save

Suppose ΔI occurs.

Then $\Delta Y = \Delta I + mpc\,(\Delta I) + mpc^2\,(\Delta I) + mpc^3\,(\Delta I) + \ldots$

$$\Delta Y = \Delta I\,(1 + mpc + mpc^2 + mpc^3 + \ldots)$$

$$\frac{\Delta Y}{\Delta I} = (1 + mpc + mpc^2 + mpc^3 + \ldots)$$

$$\frac{\Delta Y}{\Delta I} = 1 + S = \frac{1-mpc}{1-mpc} + \frac{mpc}{1-mpc} = \frac{1}{1-mpc} = \frac{1}{mps}$$

$$S = mpc + mpc^2 + mpc^3 + \ldots$$
$$S = mpc\,(1 + mpc + mpc^2 + mpc^3 \ldots)$$
$$S = mpc\,(1 + S)$$
$$S = mpc + mpc\,(S)$$
$$S\,(1 - mpc) = mpc$$
$$S = \frac{mpc}{1 - mpc}$$

<u>Example</u>: If investment rises by 100 and the mpc is .75, then the investment multiplier is 4 and GDP rises by 400.

Figure 13A.2: The Lump Sum Tax Multiplier and the Marginal Propensity to Consume

Suppose ΔT occurs.

Then $\Delta Y = -\,mpc\,(\Delta T) - mpc^2\,(\Delta T) - mpc^3\,(\Delta T) + \ldots$

$$\Delta Y = \Delta T\,(-\,mpc - mpc^2 - mpc^3 - \ldots)$$

$$\frac{\Delta Y}{\Delta T} = -\,(mpc + mpc^2 + mpc^3 + \ldots)$$

$$\frac{\Delta Y}{\Delta T} = \frac{-\,mpc}{1 - mpc}$$

Example: If a lump sum tax of 100 is imposed and the mpc is .75, then the tax multiplier is –3 and GDP falls by 300.

$$S = mpc + mpc^2 + mpc^3 + \ldots$$
$$S = mpc\,(1 + mpc + mpc^2 + mpc^3 \ldots)$$
$$S = mpc\,(1 + S)$$
$$S = mpc + mpc\,(S)$$
$$S\,(1 - mpc) = mpc$$
$$S = \frac{mpc}{1 - mpc}$$

Notes

1. See Snowdon, et al. (1994), p. 44-51, for a more advanced treatment of the classical model of output and employment.
2. See Hunt (2002), pp. 405-408, for a discussion of how Keynes's theory relates to the neoclassical flow model.
3. Hunt (2002), pp. 413-415, graphically represents the two major causes summarized here.
4. See Wolff and Resnick (2012), p. 106-107, for a discussion of how Keynes introduced this new entry point.
5. See Wolff and Resnick (2012), p. 40-41, for a discussion of the logic of Keynesian theory.
6. See Chiang and Stone (2014), p. 507-509, for the conditions under which saving ends up falling overall.
7. See U.S. Bureau of the Census (1975), Series U 187-200. Value

of Exports and Imports: 1790 to 1970, p. 884-885. The calculation of the trade balance includes total merchandise, gold, and silver.

8. Hubbard and O'Brien (2019), p. 792-793, emphasize differential growth rates across countries, differential price levels across countries, and exchange rates as the major factors influencing the level of net exports. McConnell and Brue (2008) also emphasize tariffs.
9. See Samuelson and Nordhaus (2001), p. 504-505, who make this point in the context of an example where a tax increase is required to balance the government budget after a government spending increase.
10. Chiang and Stone (2014), p. 524, and McConnell and Brue (2008), p. 188, argue that neither the income effect nor the substitution effect can explain the downward slope of the AD curve. See also Case, Fair, and Oster, p. 550.
11. These effects are identified in nearly all neoclassical textbooks when the downward sloping AD curve is explained.
12. OpenStax (2014) refers to firms "running into limits" as the economy approaches its potential GDP.
13. Many textbooks mention changes in commodity prices and changes in nominal wages. See Krugman et al. (2014), p. 422.
14. Hubbard and O'Brien (2019), p. 834, emphasize changes in the labor force and capital stock. Other books such as OpenStax (2014) describe these changes as supply shocks.
15. Prices of imported inputs are also frequently mentioned in textbooks. See OpenStax (2014), p. 562, and Samuelson and Nordhaus, p. 662.
16. See OpenStax (2014), p. 560-561, and McConnell and Brue (2008), p. 195, for a discussion of how productivity changes can influence the AS curve.
17. See Chiang and Stone (2014), p. 532.

18. See Chiang and Stone (2014), p. 531-532.
19. McConnell and Brue (2008), p. 198, analyze the post-WWII recessions in this way. McConnell and Brue (2008), p. 198, and OpenStax (2014), p. 588, provide similar graphs representing this scenario.
20. McConnell and Brue (2008), p. 199, include such arguments. Similar arguments may also be found in OpenStax (2014), p. 585. OpenStax (2014), p. 585, also mentions a coordination argument that Keynes made. That is, although workers might accept a wage cut if all workers simultaneously received one, coordinating an economy-wide wage cut is not possible in a decentralized market economy.
21. The case of demand-pull inflation is a common one in neoclassical textbooks. See Chiang and Stone (2014), p. 536-537, Samuelson and Nordhaus (2001), p. 425-426, and Bade and Parkin (2013), p. 760.
22. See Coppock and Mateer (2014), p. 436, for a similar representation of this case.
23. The case of cost-push inflation is also a common one in neoclassical textbooks. See Chiang and Stone (2014), p. 537-539, Samuelson and Nordhaus (2001), p. 426-427, and Bade and Parkin (2013), p. 761.
24. McConnell and Brue (2008), p. 200-202, analyze this case. It is also included as an exercise in Krugman et al. (2014), p. 445. Hubbard and O'Brien (2019) analyze the case too but without emphasis on the low inflation rates.
25. McConnell and Brue (2008), p. 202, make this connection to the "New Economy." Interestingly, history has repeated itself. In the 1920s, many respected observers claimed that a "new economics" had abolished the business cycle thanks to the establishment of the Federal Reserve in 1913. See Chancellor (1999), p. 192.

26. Morath, Eric. "5 Things to Watch in the Second-Quarter GDP Report." *The Wall Street Journal*. July 28, 2016. Web. https://blogs.wsj.com/briefly/2016/07/28/5-things-to-watch-in-the-first-quarter-gdp-report-3/

Chapter 14

UNORTHODOX THEORIES OF MACROECONOMIC CRISIS

Goals and Objectives:

In this chapter, we will do the following:

1. *Incorporate* turnover time into Marx's theory of competitive profit rate formation
2. *Investigate* the Marxian theory of the long-term tendency of the rate of profit to fall
3. *Analyze* the Marxian theory of the business cycle and the industrial reserve army
4. *Explore* the Marxian theory of discoordination across macroeconomic sectors
5. *Study* the causes of the 2007-2009 economic crisis from a Marxian perspective
6. *Evaluate* U.S. economic history through the lens of social structure of accumulation theory
7. *Inspect* the Austrian theory of the business cycle
8. *Contrast* Post-Keynesian effective demand theory with the neoclassical synthesis model

In Chapter 13, we investigated the neoclassical synthesis model, which represents a synthesis of neoclassical theory and Keynesian theory. The neoclassical synthesis model makes it possible to retain the neoclassical conclusion that market capitalist economies tend to return to the full employment level of output in the long even as it allows for the Keynesian conclusion that the economy may suffer from periods of depression or inflationary boom in the short run. The unorthodox theories of macroeconomic crisis that we explore in this chapter reject the neoclassical synthesis model. All assert that the tendency towards depression and periods of prolonged crisis exists in capitalist societies, but the reasons for their assertions range from the central bank's manipulation of the money supply to institutional breakdown. To explore these competing theories, we will first look at the Marxian theory of capitalist crises, which has several dimensions. After this analysis is complete, we will analyze the causes of the 2007-2009 economic crisis from a Marxian perspective. The next theory that we will consider, known as **social structure of accumulation (SSA) theory**, is a framework that many radical political economists use. It offers an original way of interpreting the history of capitalist societies and the different factors that promote capital accumulation and produce economic crises. We will then shift gears and consider the Austrian theory of the business cycle, which places most of the blame for economic crises within capitalism on the meddling of the central bank. We will conclude with a discussion of the way in which the Post-Keynesian theory of effective demand contrasts with the theory of effective demand represented in the neoclassical synthesis model.

Incorporating Turnover Time into Marx's Theory of Competitive Profit Rate Formation

In Chapter 8, we investigated Marx's theory of the formation of a competitive rate of profit. In that chapter, we considered an economy with five industries. This example has been reproduced in Table 14.1.

Table 14.1: Marx's Theory of Competitive Profit Rate Formation

	Industry 1	Industry 2	Industry 3	Industry 4	Industry 5	Entire Economy
Constant Capital ($)	80	70	50	30	20	250
Variable Capital ($)	20	30	50	70	80	250
Capital ($)	100	100	100	100	100	500
Organic Composition of Capital	80.00%	70.00%	50.00%	30.00%	20.00%	50.00%
Rate of Surplus Value	75.00%	75.00%	75.00%	75.00%	75.00%	75.00%
Surplus Value ($)	15.00	22.50	37.50	52.50	60.00	187.50
Rate of Profit	15.00%	22.50%	37.50%	52.50%	60.00%	37.50%
Value of Commodities ($)	115.00	122.50	137.50	152.50	160.00	687.50
Profit ($)	37.50	37.50	37.50	37.50	37.50	187.50
Price of Production of Commodities ($)	137.50	137.50	137.50	137.50	137.50	687.50

Because each industry has a different **organic composition of capital** (OCC) (i.e., a different ratio of constant capital to total capital), the rates of profit differ. The industries with higher rates of profit have lower organic compositions of capital. The industries with lower rates of profit have higher organic compositions of capital. It was previously explained that the industries with higher profit rates employ relatively more variable capital because labor-power is the source of value and surplus value. Because capital has a strong tendency to flow out of industries with low rates of profit and into industries with high rates of profit, the rate of profit tends to equalize across all industries as prices fall in the industries with high profit rates and prices rise in industries with low profit rates. The uniform, **general**

rate of profit (r) in this example is calculated in the following way, where S, V, and C refer to aggregate surplus value, aggregate variable capital, and aggregate constant capital, respectively:

$$r = \frac{S}{C+V} = \frac{187.50}{250+250} = 37.5\%$$

These adjustments cause **prices of production** to diverge from values and profits to diverge from surplus value in specific industries. For the overall economy, however, aggregate production price and aggregate value are equal. Additionally, aggregate profit and aggregate surplus value are equal.

The situation becomes more complicated when we introduce variations in **turnover times** across industries. In some industries, the capital turns over very quickly. That is, capital is advanced, which is to say that it is used to purchase labor-power and the means of production. The elements of production are then used to produce commodities. The commodities are then quickly sold. The quicker this transformation from money capital back into money capital occurs, the shorter the turnover time. Also, a shorter turnover time implies a greater number of turnovers per year. When the number of turnovers per year is larger, then more surplus value will be produced and appropriated in a year, other factors the same. Also, the capital is only advanced one time and is then used repeatedly throughout the year as the capital value returns to the capitalist.

The fact that capital turns over multiple times in a year leads us to a new definition of the rate of profit and a new definition of the rate of surplus value. We shall refer to these new measures as the **annual rate of surplus**

value and the **annual rate of profit**. They are defined as follows:

$$Annual\ Rate\ of\ Surplus\ Value = \frac{Annual\ Mass\ of\ Surplus\ Value}{Total\ Variable\ Capital\ Advanced}$$

$$Annual\ Rate\ of\ Profit = \frac{Annual\ Mass\ of\ Surplus\ Value}{Total\ Capital\ Advanced}$$

To calculate these measures, we need to consider the length of the turnover period in weeks, denoted as T. For example, if T = 12, then it takes 12 weeks for capital to be advanced, employed in production, and then realized through the process of exchange. We also assume that x_c and x_v represent the weekly constant capital advanced and the weekly variable capital advanced, respectively. It is essential to understand that capital must be advanced during each week of the initial turnover period. Otherwise, it would be impossible to maintain continuous production. Once the turnover period ends, the realization of commodity values guarantees that the capitalist enterprise has sufficient capital again to continue production. Therefore, the capital advanced will be equal to the product of the turnover time and the weekly capital advanced, and the annual rate of profit (r_A) will be equal to the **annual mass of surplus value** (S) divided by this capital advanced.

$$r_A = \frac{S}{T(x_c + x_v)}$$

Similarly, the annual rate of surplus value (e_A) is equal to the annual mass of surplus value divided by the variable capital advanced:

$$e_A = \frac{S}{Tx_v}$$

Using these definitions, let's consider an example of an economy with 7 industries as shown in Table 14.2.

Table 14.2: Marx's Theory of Competitive Profit Rate Formation with Variation of Turnover Time Across Industries

Circulating capital advanced per week: xc+xv	# of weeks in the turnover period: T	Total circulating capital advanced: T(xc+xv)	Real rate of surplus value: xs/xv	Surplus value produced weekly: xs	Annual mass of surplus value: S=52xs	Annual rate of surplus value: S/Txv	Annual rate of profit: S/T(xc+xv)	Annual value of commodities: 52(xc+xv+xs)	Annual production price: 52(xc+xv)+rT(xc+xv)	Annual profit	Annual rate of profit	Divergence of price from value
1. 80+20	12	1200	100%	20	1040	433%	87%	6240	7106.67	1906.67	159%	866.67
2. 70+30	10	1000	100%	30	1560	520%	156%	6760	6788.89	1588.89	159%	28.89
3. 60+40	11	1100	100%	40	2080	473%	189%	7280	6947.78	1747.78	159%	-322.22
4. 85+15	13	1300	100%	15	780	400%	60%	5980	7265.56	2065.56	159%	1285.56
5. 95+5	8	800	100%	5	260	650%	33%	5460	6471.11	1271.11	159%	1011.11
6. 50+50	9	900	100%	50	2600	578%	289%	7800	6630	1430.00	159%	-1170
7. 40+60	9	900	100%	60	3120	578%	347%	8320	6630	1430.00	159%	-1690
Total: 480+220	--	7,200	--	220	11,440	--	--	47,840	47,840	11,440	--	0

To calculate the **general annual rate of profit (r_A*)**, we need to add up the annual mass of surplus value across all industries and then divide by the total capital advanced across all n industries as follows:

$$r^*_A = \frac{\Sigma_{i=1}^{n} S_i}{\Sigma_{i=1}^{n} T_i(x_{ci}+x_{vi})}$$

When calculating the general annual rate of profit, we add up the annual surplus value across each industry i and then divide by the sum of the capital advanced across each industry i. It is this formula that is used to calculate the annual rate of profit of 159% (after rounding) in Table 14.2. That profit rate may then be used to determine the annual profit in each industry and the annual production price in each industry.

To calculate the **general annual rate of surplus value (e_A*)**, we need to add up the annual mass of surplus value across all industries and then divide by the total variable capital advanced across all n industries as follows:

$$e_A^* = \frac{\Sigma_{i=1}^n S_i}{\Sigma_{i=1}^n T_i x_{vi}}$$

It should be clear from both definitions that the annual rates of profit and surplus value are significantly higher than their daily counterparts. The reason is that capital is only advanced during the turnover period, which is typically shorter than one year, while the surplus value is appropriated throughout the entire year.

Table 14.2 reveals that the same aggregate equalities hold in an economy where different turnover periods apply to different industries. That is, the aggregate annual mass of surplus value is equal to the aggregate annual mass of profit. Also, the aggregate annual value of commodities equals the aggregate annual production price. Prior to the transformation of values into production prices, the industries with relatively more variable capital tend to have higher annual rates of profit, such as industries 6 and 7. An additional reason exists for the high annual profit rates of industries 6 and 7, namely the short turnover times of only 9 weeks in those industries. Industry 5 has the shortest turnover time of 8 weeks, but its organic composition of capital is so high that the combination produces the lowest annual rate of profit. Industry 4's low annual rate of profit can be attributed both to its high organic composition of capital and its long turnover period. Industries 1-3 have organic compositions of capital that are closer to industries 6 and 7 (which raises their annual profit rates above those of industries 4 and 5), but their long turnover periods produce lower annual profit rates than in industries 6 and 7.

This analysis effectively incorporates the turnover process into the method of transforming values into prices of production. It does not address the transformation problem, described in Chapter 8, however, and so the importance of addressing that problem should be kept in mind.

The Marxian Theory of the Long-Term Tendency of the General Rate of Profit to Fall

Marxian economists argue that the general rate of profit has a long-term tendency to fall in capitalist economies. It is a claim about the movement of the general rate of profit over a period of decades and even centuries. The explanation concentrates on capitalist competition and the way that it leads to innovation. The introduction of more advanced machinery and equipment in the production process causes an increase over time in the amount of constant capital employed in production relative to the variable capital employed in production. Because variable capital is the source of value and surplus value, the relative decline in its employment causes the general rate of profit to fall over time.

To see how an increase in the organic composition of capital tends to drive down the general rate of profit, consider the general rate of profit as we defined it before incorporating turnover time into the analysis:

$$r = \frac{S}{C+V}$$

This definition of the general rate of profit is calculated using the aggregate surplus value, the aggregate constant capital, and the aggregate variable capital. The organic composition of capital (OCC), as defined in Chapter 8, is the following:

$$OCC = \frac{C}{C+V}$$

An alternative measure of the organic composition of capital is expressed more simply as the ratio of constant capital to variable capital as follows:

$$OCC' = \frac{C}{V}$$

It is possible to rewrite the general rate of profit so that the relationship to the organic composition of capital becomes clear:

$$r = \frac{\frac{S}{V}}{\frac{C}{V}+1}$$

Other factors the same, as the organic composition of capital (OCC') increases, the general rate of profit must fall. This argument provides the explanation for the long-term tendency of the rate of profit to fall. Capitalist competition leads to innovation and a rising organic composition of capital. The general rate of profit thus tends to fall over long periods of time. The fall in the rate of profit means that capitalist enterprises have an increasingly difficult time making interest payments and rent payments out of their profits, which generates capitalist instability and economic crises with workers thrown out of work and businesses failing.

The same argument applies to the general annual rate of profit as shown below:

$$r_A^* = \frac{\Sigma_{i=1}^n S_i}{\Sigma_{i=1}^n T_i(x_{ci}+x_{vi})} = \frac{\Sigma_{i=1}^n S_i}{\Sigma_{i=1}^n T_i x_{ci}+\Sigma_{i=1}^n T_i x_{vi}} = \frac{S}{C+V} = \frac{\frac{S}{V}}{\frac{C}{V}+1}$$

The only difference here is that S refers to the annual mass of surplus value, and C and V refer to the annual aggregate constant capital advanced and the annual aggregate variable capital advanced. Both C and V are calculated according to the amounts required to maintain continuous production throughout the turnover period. The annual rate of profit also shows a tendency to decline as the organic composition of capital rises.

Although Marxian economists argue that the annual rate of profit tends to fall over long periods of time, the **law of the tendency of the rate of profit to fall (LTRPF)** is not an unconditional tendency. That is, Marxian economists, following Marx, argue that several countertendencies operate to prevent a decline of the general rate of profit. Therefore, if we observe an increase in the average rate of profit, then such movements do not subvert the law of the tendency of the rate of profit to fall. They simply mean that the countertendencies are at work and are giving a boost to the general rate of profit. We will briefly summarize the six major countertendencies that Marx identified in volume 3 of *Capital* and then add an additional countertendency to the list based on our definition of the annual rate of profit.[1]

The first factor that serves to counteract the fall in the general rate of profit is a rise in the rate of surplus value. If workers are exploited more, then S/V will rise. Other factors the same, a rise in the rate of surplus value throughout the economy will raise the general rate of profit. Therefore, even if the organic composition of capital rises due to competition and innovation, a sufficiently large increase in the rate of surplus value will nevertheless increase the general rate of profit. A likely cause of such an increase in the rate of surplus value is an extension of the length of the working day. If workers are required to work longer hours for the same wages, then the increase in absolute surplus value raises the rate of surplus value and the rate of profit. Another possible source of a rise in the rate of surplus value is a rise in productivity in the sectors that produce the means of subsistence for workers. If productivity rises in those sectors, then the value of labor-power declines with the fall in the values of the commodities that workers require daily to reproduce their labor-power. This change represents a rise in relative surplus value production. The increase in the rate of surplus value then raises the rate of profit.

A second counteracting factor that pushes back against the long-term fall in the general rate of profit is a reduction in wages below their value. When wages are pushed down below the value of labor-power, the degree of exploitation rises. The organic composition of capital rises as well, but the impact is greater on the numerator in our rewritten definition of the profit rate, which causes the rate of profit to rise overall. It is easier to see that a fall in wages below the value of labor-power will raise the rate of profit by considering the original definition (i.e., $r = S/(C+V)$). When the wages paid fall, the consequence is an unambiguous rise in the rate of profit. Different factors may cause such reductions in wages, including a weakening of unions and a period of intense competition among workers in the market for labor-power. These factors make it possible to exploit workers more, and the profit rate rises.

A third factor that serves as a countertendency to the falling rate of profit tendency is the cheapening of the elements of constant capital. Marx explained that a rise in productivity tends to increase the material elements used in the production process, but it also tends to reduce the value of those same elements. If the devaluation occurs relatively more than the increase in the material elements, then the constant capital employed will decline and the rate of profit will increase. The technological innovation that capitalist competition drives tends to raise the organic composition of capital, but the cheapening of the elements of constant capital works in the opposite direction. The organic composition of capital ends up falling, which raises the general rate of profit.

The fourth factor also involves a reduction of wages but this time it has a very specific cause. Marx referred to this factor as the relative surplus population. As capitalist development progresses and technological change advances, the introduction of labor-saving machinery leads to unemployment in many sectors of the economy. This surplus population puts downward pressure on wages, which raises the rate of exploitation. Even though it also pushes up the organic composition of capital, the impact on the numerator in our rewritten definition of the profit rate is more significant, which drives up the rate of profit overall. The same technological advances that tend to increase the organic composition of capital also drive down wages, which boosts the rate of surplus value.

A fifth factor that Marx mentions as responsible for counteracting the long-term tendency of the rate of profit to fall is international trade. That is, increased trade with foreign nations leads to imports of the elements of constant capital at lower prices. The reduction in the constant capital advanced reduces the organic composition of capital, which boosts the general rate of profit. As capitalism deepens, world trade expands, and this effect should become stronger. A similar pattern is expected with respect to the commodities that workers purchase for consumption. World trade makes it possible to import cheaper means of subsistence. As the value of labor-power declines with the increased availability of cheaper elements of consumption, the variable capital that capitalists must advance declines. The consequence is a rise in the rate of surplus value and an increase in the general rate of profit.

The sixth and final factor that Marx mentions as an offsetting factor that counteracts the long-term fall in the general rate of profit is the rise in the amount of share capital invested in production. Marx has in mind interest-bearing capital that earns interest only and thus takes a share of the average profit appropriated in industry. Moneylending capital is thus excluded from the calculation of the general rate of profit. Because it earns a rate that is far below the average rate of profit and it is frequently invested in industries with a high organic composition of capital (e.g., railroads), the consequence would be a major reduction in the general rate of profit, if it was included in the calculation of the general rate of profit. Its exclusion tends to increase the general rate of profit and thus the rise in share capital with the development of capitalism qualifies as a counteracting factor that works against the long-term tendency of the rate of profit to fall.

A final factor that we might add to Marx's list of counteracting tendencies is a change in the average turnover time across the different branches of production. The calculation of the annual rate of profit (r_A*) shows that if the turnover time in any one industry increases, then the annual rate of profit will fall. The reduction in the annual rate of profit will be even larger if most or all the industries experience longer turnover times. Similarly, if the turnover time in any one industry declines, then the annual rate of profit will rise. The rise in the annual rate of profit will be even larger when most or all industries experience shorter turnover times. Now consider what has happened to the average turnover time throughout the history of capitalism. The turnover time includes buying time (i.e., the purchase of materials and instruments of labor), production time, and selling time (i.e., the sale of the final commodities). Capitalists have long been engaged in an intense competitive struggle to appropriate more profits than competitors. A reduction in the turnover time is a primary method of reducing the capital advanced and increasing the annual mass of surplus value that is appropriated, which are factors that increase the annual rate of profit. The enormous improvements in transportation and communication technology throughout the history of capitalism have allowed capitalists to achieve this reduction of turnover time. Commodities are purchased more quickly for use in production, production itself has become immensely quicker and more efficient, and commodities are transported to the final consumer more quickly and easily than ever before. The reduction in turnover time has thus boosted the annual rate of profit and has helped counteract the long-term tendency of the annual rate of profit to fall.

Although these counteracting factors tend to increase the general rate of profit, Marxian economists assert that the long-term tendency of the profit rate to fall will reassert itself, producing economic crises, rising unemployment, and falling production. Marxian economists have more to say about the economic instability that capitalism produces, however, and so we now turn to factors that may produce economic downturns in the short-term.

A Marxian Theory of the Business Cycle and the Industrial Reserve Army

When profits are appropriated, capitalists must decide how to use them. One option is to consume it all, spending it on luxury commodities like mansions, expensive automobiles, vacations, jewelry, artwork, etc. Another option is to reinvest it to expand production. When profits are reinvested in production, Marxists state that capital is accumulated. In fact, the profits are being transformed into new capital and so the capital value grows. Capitalists can also choose to hoard the profits, but doing so will not allow it to be used for luxury consumption or profit-making and it will lose its value over time if inflation occurs.

Although these different options are available to capitalists, they are driven to accumulate capital. The intense competition that occurs among capitalists leads to capital accumulation as capitalists seek to outperform their competitors. This drive to accumulate capital has an impact on the general rate of profit, which creates the economic fluctuations that are referred to as the business cycle. To understand the reason, consider how money wages are likely to change over the course of the business cycle as depicted in Figure 14.1.[2]

Figure 14.1: The Changing Distribution of Wages and Profits Over the Course of the Business Cycle

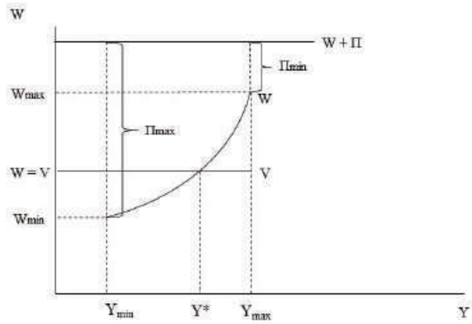


Figure 14.1 shows the value of labor-power and thus the variable capital (V) to be stable. The money wage, however, is not stable and responds to the expansion and contraction of the industrial reserve army of the unemployed. The **industrial reserve army of the unemployed** is the name that Marxian economists give to the large pool of unemployed workers that always exists within capitalist societies. It might expand or contract at different times, but it is a continuous feature of market capitalist economies. During periods of economic expansion, the reserve army of the unemployed decreases because capitalists are aggressive in their hiring. The aggressive hiring creates a shortage of labor-power and wages begin to rise above the value of labor-power. The aggregate money wages (W) thus rise above the variable capital (V), which is the value that would be advanced if the wage paid was equal to the value of labor-power. During periods of economic contraction, on the other hand, the reserve army of the unemployed increases

because capitalists dismiss workers as production slows. The rapid dismissal of workers creates a large surplus of labor-power and wages fall below the value of labor-power. The aggregate money wages spent now fall below the variable capital advanced, which again is the amount that would be spent on wages if the wage equaled the value of labor-power.

Figure 14.1 shows that as aggregate production (Y) rises from its minimum value (Y_{min}) to its maximum value (Y_{max}), the depletion of the industrial reserve army causes wages to rise from a minimum value (W_{min}) to a maximum value (W_{max}). The minimum wage fund is determined according to the minimum biological requirements to sustain the life of the worker. The value of labor-power possesses a "historical and moral element" as Marx argued.[3] Therefore, it allows a worker to purchase a basket of consumption goods that the worker needs according to a cultural norm that is specific to that society at that precise point in history. Wages may thus fall below the value of labor-power temporarily while allowing the worker to sustain his or her life and the lives of his or her dependents. At some point, the funds advanced as wages will be too low to sustain even biological life, and that amount has been designated as the lower limit.

The maximum wages are determined according to how much profits (Π) may be squeezed before it becomes impossible for capitalists to pay interest on loans, pay rent to landowners, etc. Some capitalists are in a stronger position than other capitalists and so as that point approaches, many weaker businesses fail, production begins to decline, and unemployment soars.

With the expansion of the reserve army of the unemployed, wages begin to fall and the total funds advanced as wages decline. With the decline in wages, profits start to increase. Eventually, wages fall enough and profits rise enough that capital accumulation resumes. This resumption of capital accumulation does not occur until the trough of the business cycle is reached. That is, production reaches its minimum level before capitalists can justify accumulating capital again. At that point, wages have fallen so much and capital assets have depreciated so much that new investment and expanded production are expected to be profitable.

Figures 14.2 (a) and 14.2 (b) show how the wages paid (W) and the general rate of profit (r) change over time in response to the overall fluctuations in economic activity.

Figure 14.2: Fluctuations in Wages and the General Rate of Profit Over Time

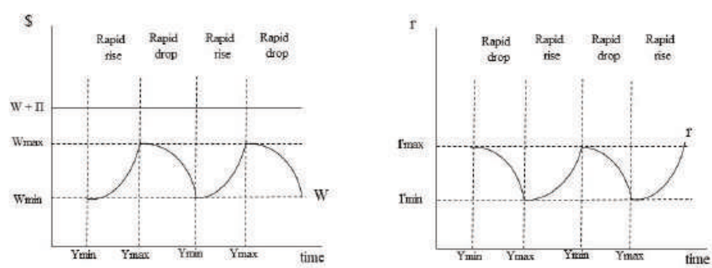


Figure 14.2 (a) Figure 14.2 (b)

Figure 14.2 (a) shows aggregate wages rising rapidly during the economic expansion and squeezing profits to a minimum level. In Figure 14.2 (a) profits are represented as the difference between aggregate wages and aggregate value added (W+Π). At the start of the economic crisis, however, unemployment soars and wages begin to drop. This reduction in wages occurs quickly and helps to restore profitability until the next expansion begins. Figure 14.2 (b) shows how the general rate of profit follows the opposite pattern relative to aggregate wages as production changes over the course of the business cycle. Because we are now thinking in terms of profits and wages as opposed to surplus value and variable capital, let's write the general rate of profit as follows:

$$r = \frac{\Pi}{C_P + W}$$

In this expression, C_P represents the production price of the means of production or the value of the constant capital transformed into its production price. We can also write the maximum value of the general rate of profit (r_{max}) and the minimum value of the general rate of profit (r_{min}) as follows:

$$r_{min} = \frac{\Pi_{min}}{C_P + W_{max}}$$

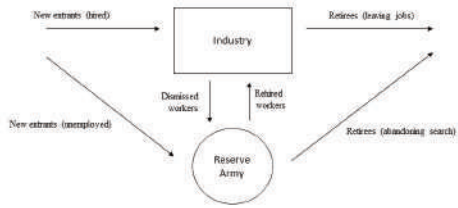
$$r_{max} = \frac{\Pi_{max}}{C_P + W_{min}}$$

As production expands, wages rise quickly, and profits are squeezed. The result is a rapid fall in the rate of profit to its minimum value, which precipitates the crisis. As production declines, aggregate wages quickly fall and profits expand for businesses that do not fail. Eventually, the rise in the general rate of profit to its maximum value makes renewed capital accumulation possible once more.

This theory of the business cycle also explains the fluctuations in the unemployment rate over time that we observe. The reserve army of the unemployed rises during contractions and falls during expansions. Nevertheless, a reserve army of the unemployed always exists. The reader should consider the contrast between the language used in neoclassical theory and in Marxian theory. In neoclassical theory, unemployment is recognized as inevitable, but it is referred to as **natural unemployment**. The use of the word "natural" suggests that a significant amount of unemployment should not be a concern to anyone. It is natural and beyond our conscious control. Marxian economists also regard a certain amount of unemployment in a capitalist society as a permanent feature of that economic system. The reference to a "reserve army" of the unemployed emphasizes the scale of the problem of unemployment. It also suggests that unemployed workers are under the control of the capitalist class and are only called into service as needed. Both schools of economic thought recognize the permanent nature of unemployment in capitalist societies, but their ways of interpreting that empirical fact are completely different.

In his classic 1942 work *The Theory of Capitalist Development,* Marxian economist Paul Sweezy includes a helpful flow diagram that captures the various factors that cause the reserve army of the unemployed to grow and shrink over time. Figure 14.3 is based on the diagram in Sweezy's book.[4]

Figure 14.3: Factors that Influence the Size of the Reserve Army of the Unemployed



Source: Sweezy, Paul. *The Theory of Capitalist Development: Principles of Marxian Political Economy.* Monthly Review Press: New York, 1970. Pp. 91. Re-created with the permission of Monthly Review Press.

The flows from the reserve army to industry and back again represent the expansions and contractions of the reserve army that have been discussed in this section. When wages rise and squeeze profits too much, workers are thrown out of work and the reserve army swells. When wages fall and profits expand, hiring picks up and the reserve army shrinks. Sweezy's model identifies additional factors that can influence the size of the reserve army of the unemployed, namely the entrance of new workers into the market for labor-power and retirements that cause workers to exit the market for labor-power. Many factors are responsible for these expansions and contractions of the reserve army of the unemployed, and our purpose here is only to demonstrate the linkage between changes in wage levels and the amount of unemployment as aggregate production rises and falls.

The Marxian Theory of Discoordination across Macroeconomic Sectors

Marx had a great deal to say about what causes capitalist crises. In volume 1 of *Capital,* Marx argued that the possibility of a crisis within capitalism was inherent within the sphere of simple commodity circulation. The reader should recall the general form of a **simple commodity circuit**, which is denoted in the following way:

$$C - M - C'$$

This circuit indicates that a commodity is sold for money and that the money is then used to purchase another commodity. The first commodity is thus transformed into a qualitatively different commodity. Classical economists, like J.B. Say, argued that crises within capitalism are impossible because "supply creates its own demand." That is, the act of selling is also an act of purchase. When commodities are sold to obtain money, the purpose is to purchase another commodity and so the act of supplying commodities also represents a demand for other commodities. Therefore, supply and demand will be equal at the aggregate level and economic crises should never occur. This argument has been labeled **Say's Law of Markets**, as we discussed in Chapter 13. The problem, as Marx pointed out, is that just because one commodity is sold for money does not mean that the money must immediately be spent on another commodity. If the seller of a commodity decides to hold on to the money, then no new demand is created. Therefore, Marx argued that the simple circulation of commodities contains within it the possibility that a crisis will occur.[5]

To argue that an event is possible is not the same as arguing that it will occur. Therefore, Marx provided additional arguments to show that capitalist crises are inevitable within capitalist societies. We have already considered Marx's argument that a rising organic composition of capital leads to a long-term decline in the general rate of profit and how this decline produces economic crises. We have also explored Marx's claim that capital accumulation leads to rising wage rates and a squeezing of profits until an economic crisis occurs, which restores wage rates to a level that is compatible with renewed capital accumulation. In this section, we will consider another aspect of Marx's theory of economic crisis that focuses on imbalances that arise across major sectors of the economy, which ultimately produce an economic crisis. Marx makes this argument using a two-sector model of social reproduction that is found in volume 2 of *Capital*.[6] The rest of this section summarizes a portion of Marx's analysis.

Table 14.3 provides information about an economy with two major sectors.

Table 14.3: The Marxian Perspective: The Theory of Capitalist Crises

Sector	Value of Means of Production	Value of Means of Consumption	Surplus Value	Total Value
I Means of Production	C_1 = \$1000	V_1 = \$500	S_1 = \$200	\$1700
II Means of Consumption	C_2 = \$700	V_2 = \$300	S_2 = \$300	\$1300
Total	C = \$1700	V = \$800	S = \$500	\$3000

In Table 14.3, one sector produces means of production and another sector produces means of consumption. The **means of consumption** include both the means of subsistence for workers and luxury commodities for capitalists. **Simple reproduction** characterizes this economy. That is, capitalists do not reinvest the surplus value to expand production and so no capital is accumulated. To see that simple reproduction exists in this economy, consider the total constant capital advanced (C). The $1700 of constant capital that is advanced is exactly equal to the value of the means of production produced in Sector I. That is, the demand for means of production exactly equals the supply of means of production. Now consider the total variable capital advanced (V) and the total surplus value (S) realized from the sale of the total social product. The $800 of wages paid to workers plus the $500 of surplus value that capitalists appropriate is just sufficient to purchase the entire $1300 of means of consumption produced in Sector II. That is, the demand for means of consumption is exactly equal to the supply of means of consumption. Since aggregate supply and aggregate demand are the same, and all surplus value is spent on luxury commodities, simple reproduction exists in this economy.

Another way to see that simple reproduction holds is to identify the condition for simple reproduction that Marx identified in this two-sector model. For simple reproduction to hold, the following condition must be met:

$$C_2 = V_1 + S_1$$

In words, the constant capital employed in Sector 2 must equal the sum of variable capital and surplus value in sector 1. We can think through the condition in the following way. Capitalists in sector 1 advance \$1000 of constant capital and so purchase \$1000 worth of means of production produced in their industry. That purchase leaves \$700 (= V_1+S_1) worth of means of production to be sold. For simple reproduction to hold, the constant capital advanced in Sector 2 must equal \$700. Otherwise, the demand for the total output of Sector 1 will be too large or too small, and an economic crisis will occur.

Alternatively, workers and capitalists in Sector 2 purchase \$600 (= V_2+S_2) worth of the means of consumption produced in their industry. That purchase leaves \$700 (= C_2) worth of means of consumption to be sold. For simple reproduction to hold, the sum of the variable capital and the surplus value in Sector 1 must equal \$700. Otherwise, the demand for the total output of Sector 2 will be too large or too small and an economic crisis will occur.

The brilliance of Marx's argument can be appreciated if we think through the sources of the different types of spending. The capitalists are willing to advance the constant capital and the variable capital. The workers then spend their wages and the capitalists purchase the means of production. Therefore, they know that when the commodities are sold, they will receive enough revenue to compensate them for these capital advances. The surplus value is a different matter. Capitalists know that they will receive the surplus value when the output is sold, but the only way that they can realize the surplus value is to advance the funds themselves! They must purchase luxury commodities using their previously realized surplus value. Why would capitalists want to use funds to make purchases only to have the same amount of surplus value return to them?

The answer in Marx's theory is that the capitalists consume a surplus product in the form of luxuries produced with the surplus labor of the working class. It is true that the capitalists advance the constant capital and the variable capital and end up with revenues equal to that amount of capital advanced plus the surplus value. In that sense, money is used to make more money, but notice that the extra money to realize the surplus value originates with the capitalists themselves. It does not matter that the capitalists end up with a sum of money that is the same as before. The capitalist class appropriates the surplus labor and consumes the surplus product.

A Marxian Analysis of the 2007-2009 Economic Crisis

Richard Wolff provides a Marxian analysis of the 2007-2009 economic crisis through a Marxian lens.[7] According to Wolff, it is not changes in aggregate investment spending, tax levels, or government spending that should be the focus of efforts to explain the crisis. Instead, we should emphasize capitalism's class structure as we struggle to understand the factors that produced the worst decline of economic activity in the United States since the Great Depression.

Wolff argues that beginning in the mid-1970s, workers' average real wages stopped rising even though they had been increasing each decade since 1820.[8] Wolff points to the displacement of American workers due to the computerization of production and the transfer of production overseas as U.S. firms searched for higher profits.[9] Even during the deflation of the Great Depression, the price level declined more rapidly than money wages, which caused real wages to rise. As Wolff explains, even after real wages began to stagnate in the 1970s, workers' productivity continued to rise, which allowed the capitalists to appropriate even more of the surplus value that workers produce.

Figure 14.4 offers a graphical depiction of the pattern of real value added per worker and real wages throughout U.S. history.

Figure 14.4: The Pattern of the Real Wage and the Real Value Added Per Worker, 1820-2006

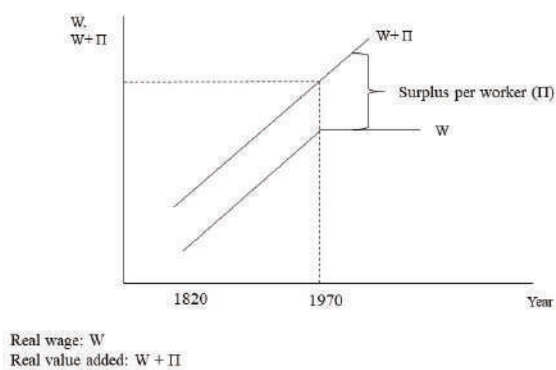


Figure 14.4 shows a gradual increase in real wages beginning in 1820 with stagnation beginning in the 1970s. Real value added per worker, on the other hand, is depicted as rising continuously up to the present time. Fluctuations are omitted from the graph to focus on the long-term trends that Wolff emphasizes. The difference

between real value added per worker and the real wage represents the surpluses extracted from workers.

Wolff explains that the surpluses were distributed in several different ways.[10] A large part of the surpluses was distributed in the form of bonuses to corporate executives. Another portion was distributed as dividends to shareholders. Yet another part was used to move production overseas. Even so, the bulk of these surpluses found their way into the banks, which transformed them into loanable money capital. The loan capital was distributed to borrowers who purchased homes and automobiles. Other borrowers paid for college tuition and consumer goods. Firms also borrowed the funds to expand production. Large financial institutions bundled together many different loans, which created specialized financial assets, such as **mortgage-backed securities (MBSs)** and **collateralized debt obligations (CDOs)**. They then sold the financial assets to large banks and institutional investors like hedge funds.

As the market for specialized financial assets exploded, so did the degree of risk in the financial system. As Wolff explains, workers began to borrow heavily to maintain a rising material standard of living.[11] Faced with stagnant wages, it was the only means of expanding upon one's material possessions. Workers thus faced a double squeeze from the 1970s to 2006, according to Wolff.[12] Capitalists took the surpluses from workers but then also took from them again as workers paid interest on their mortgage loans, auto loans, credit card loans, and student loans. Because workers had fallen so deeply into debt, they struggled to keep up with their debt payments. Many borrowers began to default on their loans. The situation was made worse because many lenders chose to lend to so-called subprime borrowers in the subprime mortgage market. **Subprime mortgage loans** are loans to people with poor credit histories and low incomes. The interest rates are high but so is the risk. Furthermore, as risky loans were made, the growth of the secondary market for the loans had expanded to the point where the originator of the loan could sell it relatively quickly. Lenders thus had less reason to be concerned about the credit worthiness of their borrowers. The more loans they pushed, the greater commission revenue they received. The incentives were thus skewed and fueled the buildup of risk within the system.

A compounding factor was the behavior of **credit rating agencies** in the leadup to the 2008 financial crisis. Standard & Poor's, Moody's, and Fitch Ratings are the three agencies most responsible for assigning ratings to financial assets like MBSs, CDOs, and other specialized financial assets. Investors rely heavily on these ratings to evaluate the degree of risk. These ratings thus affect the prices of these financial assets. When the rating assigned to an asset is high, investors infer that the degree of risk is relatively low. When the rating assigned to an asset is low, investors infer that the degree of risk is relatively high. During the housing boom prior to the Great Recession, MBSs and other assets seemed like good investments. Rising prices for these assets made them appear to be sound investments. The problem, however, is that home prices were greatly inflated, and many borrowers were taking on more debt than they could handle. The rating agencies should have recognized the high degree of risk associated with these securities and downgraded them appropriately. Because the rating agencies received fees from the large investment banks to rate the securities that they created, the rating agencies had a strong incentive to promote the growth of these financial markets. The assignment of high ratings to the newly issued securities, even when they could not be justified, furthered that goal. A **conflict of interest** exists when an individual or organization has an incentive to act in multiple, competing ways. In this case, the agencies had a mission to serve the public interest with accurate ratings provided to investors. This mission competed with its drive to maximize profits through the assignment of inflated ratings on securities. Figure 14.5 provides a diagram of the relationships that existed between investment banks, credit rating agencies, institutional investors, commercial banks, and homebuyers.

Figure 14.5: Key Linkages Contributing to the Growth of Systemic Risk in the Leadup to the 2008 Financial Crisis

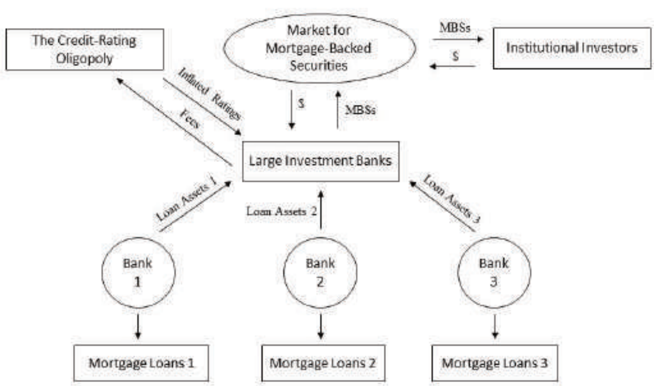


Figure 14.5 shows how commercial banks loaned money to homebuyers for the purchase of homes. The banks then sold those mortgage loan assets to investment banks,

which bundled the mortgages to create MBSs. The rating agencies then gave these securities inflated ratings to promote their further creation and sale. The newly issued securities were then sold in the financial marketplace to institutional investors such as pension funds, hedge funds, and mutual fund companies.

When home prices plummeted, many borrowers recognized that the values of their homes were below the values of their mortgage loans, and they simply walked away from their homes. Because they had put so little money down to purchase the homes, the loss of equity from a default did not stop them from abandoning their commitments. As defaults began to soar, the financial assets that represented bundles of these loans began to lose value. Large financial institutions watched as their asset values plummeted. Faced with such losses, banks and other financial institutions stopped lending to each other and to the public. Credit markets froze.[13] With firms unable to borrow money to pay wages and to purchase materials, layoffs increased, businesses failed, and unemployment soared. The Great Recession was the result of the instability in the financial system.

Wolff finds the source of the Great Recession in the class structure of capitalist society. He argues that a movement beyond a capitalist class structure would have prevented the economic crisis of 2007-2009. Specifically, he argues that workers need to become the collective appropriators of surpluses within their firms.[14] If firms were reorganized in this manner, Wolff believes that they would not have frozen real wages as firms did starting in the 1970s.[15] They also would not have distributed it as bonuses and dividends to executives and shareholders. They also would not have used it to transfer production overseas, nor would have they allowed it to accumulate in the banks to be loaned back to workers at high interest rates.[16] From a Marxian perspective, to really address the source of such instability would require revolutionary measures that challenge the right of owners to privately owned means of production.

U.S. Economic History through the Lens of Social Structure of Accumulation (SSA) Theory

A theoretical framework that is closely related to, yet distinct from, Marxian economics is social structure of accumulation (SSA) theory. This theory originated during the latter half of the twentieth century. Radical political economists use it to interpret major shifts in the history of capitalist societies. Most of the focus has been on the United States, but other national economies have been analyzed using this framework as well.

A **social structure of accumulation (SSA)** refers to an "institutional environment affecting capital accumulation."[17] During periods of stability, the set of institutions comprising a SSA promotes rapid capital accumulation and economic growth. A SSA may be in place for decades as capital accumulation and economic growth continue without major disruption. Business cycle fluctuations will occur during these long periods of time, but no major economic crisis occurs. Eventually, however, the underlying institutions that make capital accumulation possible begin to break down, and a major economic crisis occurs. The crisis creates the conditions for the restructuring of the economy. New institutions develop and eventually establish a foundation for renewed capital accumulation and economic growth.

The new institutions form a new SSA, which may do a better or worse job of promoting capital accumulation and economic growth.

This section on SSA theory concentrates on the work of Terrence McDonough who has written extensively about the nature of SSAs in U.S. economic history. [18]McDonough argues that three SSAs may be identified as we look back at the history of American capitalism: a post-Civil War SSA, a monopoly capitalist SSA, and a post-WWII SSA.[19] McDonough does not devote much attention to the **post-Civil War SSA** but argues that it represents a period of **primitive accumulation**, thus revealing the Marxian roots of SSA theory.[20] It was thus a period during which many of the resources of the nation were transformed into privately owned means of production and pulled into the capital accumulation process.

McDonough also describes the **monopoly capitalist SSA**, which took shape during the late nineteenth and early twentieth centuries. He argues that each SSA is built according to a unique organizing principle.[21] The organizing principle in the case of the monopoly capitalist SSA is the oligopolistic market structure that developed due to the wave of corporate mergers that occurred during those years.[22] A specific subset of institutions is also needed to revive the capital accumulation process after an economic crisis, which David Kotz calls the **core of the SSA**.[23] According to Kotz, the core institutions of a SSA moderate class conflict and capitalist competition during the long period of expansion.[24] During the monopoly capitalist SSA, a greater concentration of industry was one of the core institutions.[25] McDonough identifies several factors that contributed, including a growing market for industrial securities, New Jersey holding company legislation, and Sherman Act interpretations that permitted monopoly by merger.[26] A second core institution was an electoral shift towards the Republican Party at the federal level. Republican control of Congress and the Presidency in the early twentieth century led to policies that supported financial and industrial capitalists (e.g., protectionism).[27]

A third core institution that McDonough identifies with the monopoly capitalist SSA is the regulation of trusts rather than the breaking up of trusts.[28] The administration of Theodore Roosevelt would distinguish between trusts that were behaving responsibly and trusts that abused their power. It created a Bureau of Corporations in 1903 that would publicize abuses of economic power to keep large corporations in line.[29] A fourth core institution of the monopoly capitalist SSA is a new ideological system referred to as corporatism.[30] According to the corporatist ideology, cooperation among capital, labor, and the public was a worthwhile goal.[31] Through the National Civic Federation (NCF), founded in 1900, business leaders, political leaders, and labor leaders promoted cooperation among these entities.[32] This set of ideas thus helped create a stable environment for capital accumulation and economic growth.

A fifth core institution that McDonough identifies with the monopoly capitalist SSA is a change in capital-labor relations. This development consisted of two parts. The first part involved an effort to break the power that the skilled workers had over the production process. This goal was accomplished via the reduction of semi-skilled labor to a common denominator using highly mechanized production.[33] As machines began to perform more work, and the workers became more like operators of machinery than skilled craftsmen, the result was a loss of worker control over the production process. The second part of the change in capital-labor relations, according to McDonough, was the anti-union strategy that employers adopted.[34] As workers began to organize in response to their loss of control of production, employers implemented an open-shop policy,[35] which means that the employers did not recognize unions at all and would not distinguish between union members and non-union workers in the hiring process. In other words, the employer would only negotiate with individual workers rather than a collective body.

The final core institution that McDonough associates with the monopoly capitalist SSA is an imperialist strategy on the part of the United States beginning in the late 1890s. The U.S. aggressively sought to expand into foreign markets to offset the effects of the 1890s slump.[36] McDonough mentions the U.S. entry into the Spanish-American War, the annexation of Hawaii, the treatment of Cuba as a protectorate of the United States, and the annexation of the Philippines, Puerto Rico, and Guam.[37] President McKinley's "Open Door" policy towards China was also part of the U.S. effort to establish itself as a dominant player in the sphere of international trade.[38]

Eventually, the monopoly capitalist SSA experienced a period of prolonged economic crisis. The world wars and the worldwide Great Depression represented a crisis of the institutions that had served capital accumulation in the early twentieth century. From the ashes of the old SSA rose a new set of institutions that established a foundation for rapid capital accumulation and economic growth in the post-World War II period. The **postwar SSA**, as McDonough calls it, had five core institutions.[39] According to McDonough, "the social influence of the war itself" was the organizing factor for this SSA.[40]

The first core institution of the postwar SSA is the Keynesian state. The Second World War demonstrated that Keynesian policies could be effective, and Keynesian economists began to acquire government positions.[41] Full employment became a goal and with the expansion of the public sector, changes in government spending or taxes could influence aggregate output.[42] Aggregate demand received a boost due to the 1947 Marshall Plan, which raised overseas demand, and the Korean War, which increased military spending.[43]

The third and fourth core institutions of the postwar SSA include the international dominance of the United States and the adoption of a Cold War ideology that guided U.S. policymaking after World War II. According to McDonough, the U.S. became the most powerful nation economically and militarily.[44] It used the 1944 Bretton Woods agreements and the 1947 Marshall Plan to establish "a worldwide capitalist economy open to American investment and export."[45] The Truman doctrine of containment insisted on a connection between Soviet ideology and the tendency towards international expansion.[46] The Cold War ideology helped discourage the spread of socialist ideas and policies just as U.S. international military and economic dominance helped ensure the creation of a global capitalist order.

The fifth core institution of the postwar SSA involved a new relationship between capital and labor. Federal support for collective bargaining, as represented in the passage of the Wagner Act in 1935, led to the growth of membership in industrial unions. **Industrial unions** aim to organize all the workers in an industry like the automobile industry or the steel industry. The United Auto Workers (UAW) and the United Steelworkers of America (USWA) are examples of industrial unions. **Craft unions**, on the other hand, aim to organize all the workers with a specific skill, like carpenters or electrical workers. McDonough explains that during the postwar period, a rough equilibrium resulted between labor and management in which management granted automatic cost of living adjustments (COLAs) and productivity-linked wage contracts in exchange for management control of the production process.[47]

The final core institution of the postwar SSA was a shift towards the Democratic Party in national politics. The political realignment involved greater capitalist support for internationalization and a willingness to cooperate with organized labor.[48] It represented a departure from Republican support for protectionism and anti-union tactics, which dominated in the past.[49] Support for the Democrats in national elections came from labor unions, the lower-class vote, and capitalists in capital-intensive industries.[50]

Whereas the core institutions of the monopoly capitalist SSA were organized around the oligopolistic market structure of the early twentieth century, McDonough argues that the core institutions of the postwar SSA were organized around the characteristics of the war itself. He argues that even though a general principle of SSA construction appears to be at the center of each SSA, it seems to be a different principle with each SSA.[51] We are thus unable to predict the organizing principle of future SSAs or their timing.[52] Nevertheless, the power of the SSA framework lies in its ability to shed light on the historical factors that create a basis for rapid capital accumulation and economic growth. When those elements begin to unravel, the sources of widespread economic crisis also become clear when viewing capitalist history through the lens of SSA theory.

The Austrian Theory of the Business Cycle

In this section we consider a theory of the business cycle that Austrian economists developed. Ludwig von Mises and F.A. Hayek are the major contributors to this theory, although it has been refined and developed since they worked on the subject. Austrian economist Roger Garrison offers a helpful overview of the Austrian theory of the business cycle using a variety of graphs that facilitate a comparison with the Keynesian cross model.[53] This section borrows heavily from Garrison's well-known essay to introduce students to the key elements of the Austrian perspective.[54]

In their business cycle theory, Austrian economists refer to capitalists and laborers. The two groups are not in conflict, however, which marks a major difference between the Austrian perspective and the Marxian and Post-Keynesian perspectives. Drawing upon Hayek's work, Austrian economists use a graph to represent the **structure of production**. The structure of production has two key elements in Austrian theory: the quantity of capital employed and the period of production.[55] These two elements are positively related in Austrian economic theory. That is, when more capital is employed, a longer period of production becomes possible.[56] The graph that is used to represent the structure of production is referred to as a **Hayekian triangle**. An example of a Hayekian triangle, which Garrison modifies somewhat to look like a trapezoid,[57] is shown in Figure 14.6.

Figure 14.6: The Structure of Production as Captured in Garrison's Version of the Hayekian Triangle

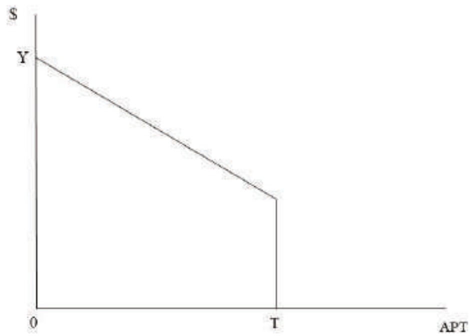


Figure 14.6 measures the aggregate production time (APT) on the horizontal axis and the dollar value of total output (Y) on the vertical axis. Beginning at point T and moving towards the origin, time passes. As time passes, the capital value increases as unfinished goods move closer to the point of completion when the consumption goods with a value of Y emerge.[58] Austrian economists regard labor and land as the **original means of production** because they have not passed through a prior production process.[59] The vertical portion in Figure 14.6 represents the value of land and labor employed in production at the start of the production period. Any increase in the value of output beyond that amount represents the contribution of capital alone and so no

exploitation of labor-power exists in the Austrian model in contrast to Marxian theory.

The slope of the line in Figure 14.6 represents the rate of interest or the rate of profit for capitalists[60] because it shows how the capital value increases for each unit of time that passes in the production period. To understand how the rate of interest is determined in Austrian theory, we must consider the market for present goods. When laborers sell their labor services, it constitutes a demand for present goods (D_{PG}) because they will spend their wages, and it represents a supply of future goods (S_{FG}) because they will produce goods to be sold in the future.[61] This relationship can be expressed symbolically as follows:

$D_{PG} \Leftrightarrow S_{FG}$

Similarly, when capitalists purchase labor services, it represents a supply of present goods (S_{PG}) because they must advance goods to laborers for present consumption and a demand for future goods (D_{FG}) because they are postponing their own consumption in the present when they pay workers.[62] This relationship may also be represented symbolically:

$S_{PG} \Leftrightarrow D_{FG}$

Following Garrison,[63] the market for present goods is represented in Figure 14.7.

Figure 14.7: The Market for Present Goods

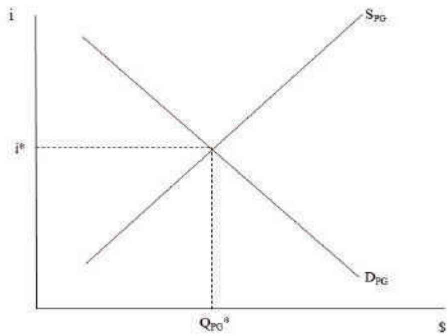
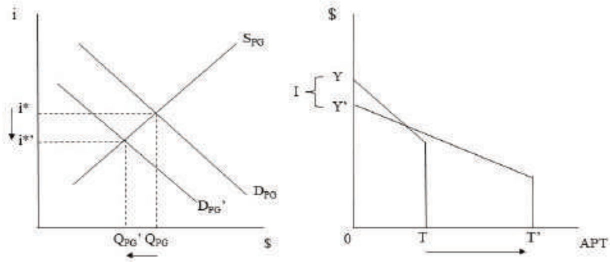


Figure 14.7 shows how an equilibrium interest rate and an equilibrium quantity exchanged of present goods is determined in the competitive market.[64] Laborers are represented on the demand side of the market, and capitalists are represented on the supply side of the market. We can think about the slopes of the curves as follows. When the rate of interest falls, laborers will prefer to save less and consume more today, which produces a downward sloping demand curve. When the interest rate rises, capitalists will anticipate greater profits by purchasing labor services and so they will supply more present goods. If the market is not in equilibrium due to a relatively high rate of interest, then the surplus of present goods implies a rise in savings that causes the rate of interest to fall. If the interest rate is below the equilibrium interest rate, then the shortage of present goods implies a reduction in saving and a rise in the rate of interest.

It is the time preferences of laborers and capitalists, however, that determine the positions of the supply and demand curves for present goods.[65] **Time preference** refers to an individual's preference to consume in the present rather than in the future. If an individual has a **high time preference**, then consumption is preferred today much more than in the future. If an individual has a **low time preference**, then consumption is preferred in the future much more than in the present. In the market for present goods, a higher time preference for laborers would lead to a rise in the demand for present goods and a rightward shift of the demand curve. A lower time preference for laborers, on the other hand, would lead to a fall in the demand for present goods and a leftward shift of the demand curve.

Now consider how a change in time preferences alters the structure of production. A reduction in the time preferences of laborers will cause the demand for present goods to decline[66] and a leftward shift of the demand curve as shown in Figure 14.8.

Figure 14.8: A Change in the Structure of Production Due to a Change in the Time Preferences of Laborers



The consequence of the decline in the demand for present goods is a reduction in the rate of interest and in the quantity of present goods exchanged in the market. Since workers demand fewer present consumption goods, investment rises as resources are released from consumer goods production.[67] Capital and labor are thus reallocated to lengthier production processes.[68] Real consumption falls temporarily but eventually rises when the new consumer goods enter the market. Consumer goods prices fall and so the dollar value of consumer goods declines even as real consumption rises.[69] Because output remains the same, we can see how consumption and investment now make up total output. Figure 14.8 shows how the structure of production has changed. The rise in investment causes the aggregate production period to lengthen. The slope of the Hayekian triangle is smaller due to the fall in the rate of interest. Finally, the value of the original means of production has fallen because fewer present goods are advanced for laborers causing the vertical portion of the Hayekian triangle (trapezoid) to be smaller.

Now consider how a change in the time preferences of capitalists alters the structure of production. A decrease in the time preferences of capitalists will cause the supply of present goods to rise and a rightward shift of the supply curve for present goods. Capitalists will advance more present goods to enjoy more real consumption later.[70] Figure 14.9 shows how a rightward shift of the supply curve affects the rate of interest and the equilibrium quantity of present goods exchanged.

Figure 14.9: A Change in the Structure of Production Due to a Change in the Time Preferences of Capitalists

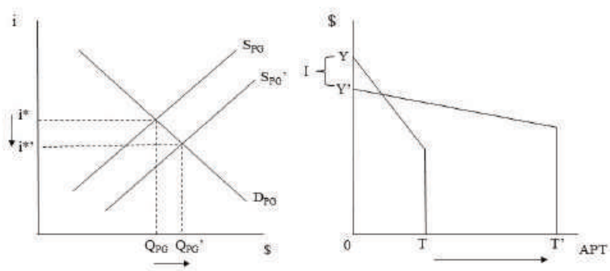


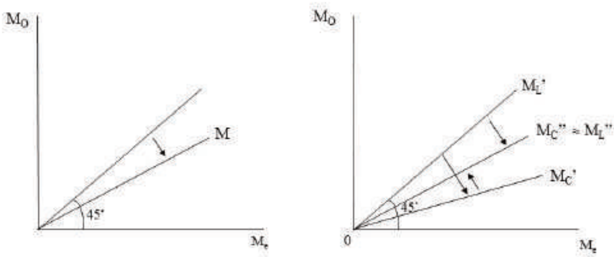
Figure 14.9 shows that the increase in the supply of present goods causes a reduction in the rate of interest and a rise in the equilibrium quantity of present goods exchanged. The result is a rise in investment as entrepreneurs observe the increased demand for future goods. More investment makes possible a longer

aggregate period of production. The fall in the rate of interest makes the slope of the Hayekian triangle smaller as well. Finally, the vertical portion of the Hayekian triangle (trapezoid) becomes larger because more present goods are advanced to laborers. As in the case of a reduction in the demand for present goods, real consumption falls but only temporarily. Once the new consumer goods reach the market, real consumption expands.

At this stage, we can see a relationship between the Keynesian cross model and the Austrian theory of the business cycle. In the Keynesian cross model, consumer spending is combined with exogenously determined investment spending to determine aggregate expenditure and ultimately the equilibrium aggregate output of the economy. In Austrian business cycle theory, on the other hand, investment spending is endogenously determined as the time preferences of capitalists and laborers change.[71] Garrison sets the Austrian graph of the structure of production next to the Keynesian Cross model to illustrate the relationships between the two economic models.[72]

Nothing in the Austrian analysis thus far suggests a cause of the business cycle. In fact, the economy will function without any major disruption as producers respond to shifts in the time preferences of laborers and capitalists. The major cause of depressions is found in central bank manipulation of the money supply. Austrian economists thus introduce the money supply into the analysis as an exogenous variable.[73] A major difference exists between Austrian monetary theory and neoclassical monetary theory. Whereas neoclassical economists argue that “new money is injected uniformly throughout the economy,” Austrian economists argue that injections of new money tend to fall into the hands of producers.[74] Figure 14.10 shows how Garrison represents neutral and non-neutral monetary expansions.[75]

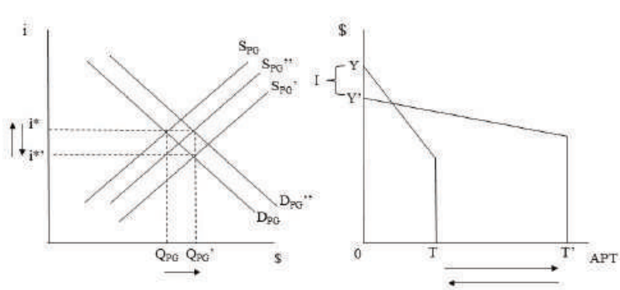
Figure 14.10: Neutral Versus Non-Neutral Monetary Expansions



The vertical axis shows the money stock before the expansion. The horizontal axis measures the money stock after the expansion. The 45-degree line is a reference line that shows us what happens when no monetary expansion occurs since the money stock is the same before and after the "expansion." The **neutral monetary expansion** assumes that all capitalists and laborers receive the new money in the same proportion. The **non-neutral monetary expansion**, on the other hand, shows that the money in the hands of capitalists expands first before it spreads throughout the economy as capitalists pay wages to laborers. Eventually, the laborers and capitalists are affected in the same way due to the monetary expansion when the lines for each group coincide.

In the case of a non-neutral monetary expansion, we can see how the structure of production is affected. First, the monetary injection leads to an increase in **producer credits** (i.e., money in the hands of capitalists). The capitalists use it to increase the supply of present goods.[76] The rightward shift of the supply curve for present goods causes the rate of interest to fall. As before, this change will alter the structure of production. It leads to an increase in investment, a lengthening of the aggregate period of production, and an increase in the value of the original means of production. As the monetary injection begins to flow throughout the economy, it causes laborers to increase their demand for present goods and the supply of present goods will begin to contract until the rate of interest returns to its original level,[77] as shown in Figure 14.11.

Figure 14.11: A Change in the Structure of Production Due to a Non-Neutral Monetary Expansion (short run and long run)



The return of the rate of interest to its original level causes the structure of production to return to its original state. The monetary expansion pushes the **market rate of interest** down below the **natural rate of interest**. It is the natural rate of interest in the Austrian model that is consistent with the time preferences of capitalists and laborers. The increase in investment causes a surge of demand for the remaining consumer goods because the time preferences of individuals have not actually changed.[78] The price of consumer goods relative to capital goods thus rises. Entrepreneurs recognize their error and begin to liquidate their investment projects.[79] The consequence is disinvestment on a large scale, and an economic crisis ensues. The source of economic crises is thus found in a misallocation of capital and what Austrian economists call **forced saving** because the shift towards investment goods does not reflect a real change in the time preferences of individuals.

Economic crises that Austrian economists explain using this theory of the business cycle include the Great Depression of the 1930s and the Great Recession of 2007-2009. The Federal Reserve is blamed for expanding the money supply and artificially pushing down the rate of interest. The result was excessive investment in the 1920s and in the early 2000s in the housing market or in the 1990s in the IT sector. Eventually, the errors of entrepreneurs became apparent and investment projects were abandoned, leading to major economic contractions. According to Snowdon et al., an **iron law of retribution** exists.[80] That is, the greater the monetary expansion and economic boom, the greater the contraction and disinvestment to follow will be. In fact, the correction to the economy might be so great that the disinvestment might lead to **capital consumption** where the economy ends up with a smaller capital stock than it possessed at the beginning of the process.[81] If the government provides **consumer credits** (i.e., cash subsidies for laborers) to pump up demand, it will boost the demand for consumer goods and alter relative prices in favor of consumer goods even more. The result will be an even larger contraction of investment projects that makes the economic crisis worse.[82]

The Austrian theory of the business cycle is noteworthy because it offers a theory of capitalist crises that assigns a central role to the difference between capitalists and laborers even as it places the blame on government intervention and central bank manipulation of the money supply. It also recognizes that entrepreneurs do not have perfect foresight and that they may make mistakes when they increase investment in response to a drop in the market rate of interest that stems from a monetary expansion. Finally, it provides an alternative to Keynesian theory when thinking about the factors that influence aggregate investment, the aggregate production period, the rate of interest, and aggregate output.

Post-Keynesian Effective Demand Theory versus the Neoclassical Synthesis Model

Like neoclassical economists, Post-Keynesian economists recognize that changes in key macroeconomic variables influence aggregate output and employment. Unlike neoclassical economists, however, Post-Keynesian economists place much emphasis on the class-based distribution of income. We may begin this discussion with the simple fact that aggregate output or aggregate

income (Y) can be decomposed into **aggregate wages (W)** and **aggregate profits (Π)** as follows:
$$Y = W + \Pi$$

When the distribution of income is represented as a division between wages and profits, it is referred to as the **functional distribution of income**.

Post-Keynesian economists agree with neoclassical economists that aggregate expenditure (A) may be written as the sum of the major spending components from the national income and product accounts.
$$A = C + I + G + X - M$$

Aggregate expenditure thus represents the sum of consumer spending (C), investment spending (I), government spending (G), and net exports (X – M). Net exports are calculated as the difference between exports (X) and imports (M).

Post-Keynesian economists also agree with the macroeconomic equilibrium condition that aggregate output (Y) equals aggregate planned expenditure (A), as shown below:
$$Y = A$$

If we substitute the expression for aggregate income and the expression for aggregate expenditure into the equilibrium condition, we obtain the following result:
$$W + \Pi = C + I + G + X - M$$

To further modify this equation, let's divide aggregate consumer spending into the consumer spending by capitalists (C_c) and the consumer spending by workers (C_W) as follows:

$$W + \Pi = C_C + C_W + I + G + X - M$$

Workers' consumption may be rewritten as what remains of workers' wages after they have paid their taxes (T) and set aside a portion to be saved (S_W).

$$C_W = W - T - S_W$$

The equilibrium condition (Y = A) may now be written as follows:

$$W + \Pi = C_C + W - T - S_W + I + G + X - M$$

Eliminating aggregate wages from both sides and rearranging the terms, we obtain the following result:[83]

$$\Pi = I + (G - T) + (X - M) + C_C - S_W$$

This result relates the **equilibrium aggregate profit** in the economy to various components of aggregate spending, such as the **government budget gap** (G – T) and the **trade balance** (X – M). That is, this equation allows us to understand how aggregate profits change over the course of the business cycle as different components of aggregate spending change. For example, if investment spending increases, then aggregate profits will rise, other factors held constant. Whereas neoclassical Keynesian economists focus on the impact of a rise in investment spending on aggregate output, Post-Keynesian economists focus on the impact on aggregate profits. Because aggregate wages do not change when investment spending rises, aggregate profits increase relative to aggregate wages. Capitalists thus gain relative to workers, and income inequality worsens.

Now consider how changes in the government budget affect profitability and the distribution of income. If government spending increases and taxes fall, then both factors create a larger government budget deficit. Deficit spending increases the equilibrium aggregate profits in the economy just as it increases the equilibrium output. With unchanged wages, the result is an increase in aggregate profits relative to aggregate wages and a rise in income inequality as the economy grows. Hence, higher government spending or tax cuts can worsen inequality even as they stimulate the economy.

If the level of net exports increases due to a rise in exports or a reduction in imports, then the result is a higher level of aggregate profits. With aggregate profits rising relative to aggregate wages, the result is worsening income inequality. At the same time, aggregate output and employment increase with rising net exports. We can conclude then that a trade surplus boosts economic growth in the short run, but it also worsens the class-based distribution of income. A trade deficit has the opposite effects. Because net exports are negative in the case of a trade deficit, the result is a lower level of aggregate output and employment. With falling profitability, the result is a reduction in income inequality.

Another component of aggregate spending that can influence the equilibrium level of aggregate profits is capitalist consumption. If capitalists increase their aggregate consumption level, then the economy receives a boost. The rise in aggregate consumption increases equilibrium output but also aggregate profits. The consequence is worsening income inequality because aggregate profits rise relative to aggregate wages. Here we can see the Marxian roots of Post-Keynesian economics at play. In Marx's two-sector macroeconomic model discussed earlier in this chapter, capitalist consumption makes possible the realization of surplus value in both sectors. Similarly, in this Post-Keynesian model, capitalist consumption enhances the profitability of capitalists. Also, in Marxian economics, workers are paid according to the value of labor-power. If aggregate wages are determined according to the value of aggregate labor-power, then wages will be relatively stable. Rising profits and aggregate income then tend to worsen income inequality over the course of the business cycle. The connection to Marxian economics is not a perfect one, however, because Marxists also argue that wages tend to rise and fall over the course of the business cycle as wages rise above and fall below the value of labor-power.

Finally, a rise in workers' savings will influence the equilibrium aggregate profits in the economy. For example, if workers save more, then aggregate demand falls, which reduces the equilibrium aggregate output and the equilibrium aggregate profits. Income inequality lessens due to increased saving on the part of workers as aggregate profits fall relative to aggregate wages.

The major difference between this Post-Keynesian approach to macroeconomic fluctuations and the neoclassical synthesis model is that neoclassical economists only focus on changes in the equilibrium level of output over the course of the business cycle. Post-Keynesian economists, on the other hand, also emphasize changes in profitability and the distribution of income as the economy experiences economic booms and busts.

Following the Economic News[84]

More than 25 million people applied online for approximately 90,000 positions in India's government-run railway system as of March 2018, according to Neha Dasgupta. The Indian Prime Minister, Narendra Modi, has set a goal of increasing the percentage of output for which manufacturing accounts in the Indian economy from 17% to 25% and to create 100 million jobs by 2022. The railway has been hiring "engine drivers, technicians, carpenters, track inspection crews and other roles related to improving safety in the world's fourth-largest network." Dasgupta explains that more than 1 million young people are entering the labor force each month in search of jobs. According to one CEO, the enormous interest in these positions reveals the considerable stress that unemployed workers are experiencing. Written tests and physical fitness tests will be used to determine who is hired. Dasgupta also explains that India's unemployment rate hit a 15-month high in February 2018 at 6.1%.

The high Indian unemployment rate is a perfect example of Marx's reserve army of the unemployed. The huge surplus of available labor-power helps keep wages low in the market for labor-power and profits high. The accumulation of capital that this situation creates may ultimately cause unemployment to fall and wages to rise, which could produce a macroeconomic crisis. Using the Post-Keynesian model of effective demand, we can also see that the railway expansion is likely to increase the equilibrium aggregate profits in the economy given the effect on government spending. An expansion of the economy is likely to occur at least in the short-term as well as an increase in income inequality as aggregate profits rise relative to wages. Finally, Austrian business cycle theory leads us to expect a different outcome. If the

Indian central bank accommodates the large increase in government spending with a huge increase in the money supply, then the rate of interest might fall even though no changes in the time preferences of capitalists and laborers have occurred. The result is an increase in the aggregate production time and a surge of investment at the expense of consumption. As Indian consumers bid the prices of the remaining consumer goods upwards, a massive amount of disinvestment will follow and an economic collapse will occur. Depending upon one's theoretical lens, it is possible to view the same situation of high Indian unemployment from several different angles.

Summary of Key Points

1. A general rate of profit and prices of production are formed when capital flows out of industries with relatively high organic compositions of capital and into industries with relatively low organic compositions of capital.
2. Other factors held constant, a longer turnover time reduces the annual rate of profit, and a shorter turnover time increases the annual rate of profit.
3. Other factors held constant, a rise in the organic composition of capital over long periods causes the general annual rate of profit to fall, which produces economic crises.
4. Several countertendencies operate to prevent a decline in the general rate of profit, including an increase in the rate of surplus value, a reduction in wages below their value, a cheapening of the elements of constant capital, and a shortening of the turnover processes of different capitals.

5. During periods of rapid capital accumulation, money wages rise above the value of labor-power as the reserve army of the unemployed declines and the rate of profit falls, which ultimately produces an economic crisis. During periods of economic contraction, money wages fall below the value of labor-power as the reserve army of the unemployed expands and the rate of profit begins to rise, which eventually makes an economic recovery possible.
6. In Marx's two-sector model, the constant capital employed in Sector 2 must equal the sum of the variable capital and the surplus value in sector 1. Otherwise, discoordination across sectors will occur at the macroeconomic level, and an economic crisis will occur.
7. According to Richard Wolff, stagnant wages and rising productivity have made possible a major expansion of surpluses and lending since the 1970s, which set the stage for the 2008 financial crisis and the Great Recession of 2007-2009.
8. Social structure of accumulation (SSA) theory is a framework that radical political economists use to understand how human institutions contribute to long periods of rapid capital accumulation and economic growth.
9. According to Terence McDonough, the organizing principle of the monopoly capitalist SSA is the oligopolistic market structure of the early twentieth century, and the organizing principle of the postwar SSA is the social influence of the Second World War.
10. Austrian business cycle theory maintains that the structure of production depends on the

aggregate period of production and the quantity of capital employed in production.

11. Austrian business cycle theorists argue that the structure of production may change when the time preferences of capitalists and laborers change. It may also change when the central bank creates an artificial boom via monetary expansion, but in the latter case, an economic crisis will follow when entrepreneurs realize that they have misjudged the time preferences of the consuming public.
12. The Post-Keynesian theory of effective demand clarifies how the equilibrium profits of the economy and the degree of income inequality change in response to changes in the components of aggregate spending.

List of Key Terms

Social structure of accumulation (SSA) theory

Organic composition of capital (OCC)

General rate of profit (r)

Prices of production

Turnover time

Annual rate of surplus value

Annual rate of profit

Annual mass of surplus value

General annual rate of profit (r_A*)

General annual rate of surplus value (e_A*)

Law of the tendency of the rate of profit to fall (LTRPF)

Industrial reserve army of the unemployed

Natural unemployment

Simple commodity circuit

Say's Law of Markets

Means of consumption

Simple reproduction

Mortgage-backed securities (MBSs)

Collateralized debt obligations (CDOs)

Subprime mortgage loans

Credit rating agencies

Conflict of interest

Social structure of accumulation (SSA)

Post-Civil War SSA

Primitive accumulation

Monopoly capitalist SSA

Core of the SSA

Postwar SSA

Industrial unions

Craft unions

Structure of production

Hayekian triangle

Original means of production

Time preference

High time preference

Low time preference

Neutral monetary expansion

Non-neutral monetary expansion

Producer credits

Market rate of interest

Natural rate of interest

Forced saving

Iron law of retribution

Capital consumption

Consumer credits

Aggregate wages (W)

Aggregate profits (Π)

Functional distribution of income

Equilibrium aggregate profit

Government budget gap

Trade balance

Problems for Review

1. Consider an economy with only two capitalist enterprises. Assume that enterprise A advances weekly constant capital of $200 and weekly variable capital of $100. Assume that enterprise B advances weekly constant capital of $100 and weekly variable capital of $200. Also assume that enterprise A has a turnover time of 4 weeks and enterprise B has a turnover time of 8 weeks. Calculate the annual rate of profit and the annual rate of surplus value assuming the weekly rate of surplus value is 100% for both enterprises and that each year has 52 weeks. How do the annual rates of profit compare? Is this result surprising? Explain the result with reference to the organic compositions of capital and to the turnover times of the two enterprises.

2. Suppose competition among capitalists in the international commodity markets has pushed down the prices of imported raw materials that are used in many domestic industries. What is the impact on the aggregate constant capital advanced? What is the impact on the organic composition of capital at the aggregate level? What is the impact on the annual general rate of profit? How are these changes important in terms of the long-term trajectory of the domestic capitalist economy?

3. Suppose that a critic of Marxian economics argues that surplus value cannot persist within capitalist economies because competition among capitalists for the source of surplus value (i.e., labor-power) will drive up wages until all the surplus value vanishes. The critic

concludes that profits must have their source elsewhere. How can Marxian economic theory be used to refute this argument?

4. Consider an economy with two major sectors: Sector I produces the means of production, and Sector II produces the means of consumption. Suppose that the constant capital advanced in Sectors I and II are $20 billion and $30 billion, respectively. Assume that the variable capital advanced in Sectors I and II are $20 billion and $10 billion, respectively. Also assume that the rate of surplus value is 100% in each sector. Is the economy balanced or will a macroeconomic crisis occur due to discoordination between the two sectors? Explain with reference to the numerical values in this example.

5. Consider the conflict of interest that arose for the credit rating agencies leading up to the financial crisis of 2008. What do you think is the best way to limit such conflicts of interest that might arise in financial markets?

6. Create a table with two columns. In one column, list the core institutions of the monopoly SSA. In the second column, list the core institutions of the postwar SSA. Try to arrange the sets of institutions so that they correspond to each other as much as possible. Finally, compare the core institutions of the two SSAs and note how they differ and how they are similar.

7. Consider Austrian business cycle theory when answering this question. Suppose that laborers experience an increase in their time preferences. How will the market for present goods be affected? Draw the changes on a graph and explain in words what is happening. How will the structure of production be

affected? Draw the changes on a graph and explain in words what is happening. Be sure to refer to the aggregate period of production, the rate of interest, and the level of investment when answering this question.

8. Consider Post-Keynesian theory when answering this question. Suppose that workers reduce their saving. What will happen to capitalist profits and the degree of income inequality? To which phase of the business cycle does this change correspond? Is this change consistent with what you would expect to happen using the neoclassical synthesis model?

Notes

1. For Marx's analysis of the countertendencies to the law of the falling rate of profit, see Marx (1991), p. 339-348.
2. The modeling approach used in Figures 14.1 and 14.2 is inspired by the approach found in Lianos (1987). Lianos's use of this approach in the context of financial markets is discussed in Chapter 15.
3. Marx (1990), p. 275.
4. I am deeply grateful to Monthly Review Press for granting me permission to include a re-creation of Sweezy's (1970) figure, p. 91.
5. See Marx (1990), p. 209.
6. See Marx (1992), p. 471-478.
7. Wolff (2010), p. 83-86.
8. Ibid. p. 83.
9. Ibid. p. 83.

10. Ibid. p. 83.
11. Ibid. p. 84.
12. Ibid. p. 84.
13. Ibid. p. 85.
14. Ibid. p. 85.
15. Ibid. p. 85.
16. Ibid. p. 85.
17. Wolfson (1994).
18. I am deeply grateful to Cambridge University Press for granting me permission to include extensive citations from McDonough's book chapter.
19. McDonough (1994), p. 103.
20. Ibid. p. 103.
21. Ibid. p. 103.
22. Ibid. p. 104.
23. Ibid. p. 104.
24. Ibid. p. 105.
25. Ibid. p. 105.
26. Ibid. p. 106.
27. Ibid. p. 107.
28. Ibid. p. 108.
29. Ibid. p. 108.
30. Ibid. p. 108.
31. Ibid. p. 109.
32. Ibid. p. 109.
33. Ibid. p. 110.
34. Ibid. p. 110.
35. Ibid. p. 110.

36. Ibid. p. 110.
37. Ibid. p. 111.
38. Ibid. p. 111.
39. Ibid. p. 114.
40. Ibid. p. 115.
41. Ibid. p. 116.
42. Ibid. p. 116-117.
43. Ibid. p. 117.
44. Ibid. p. 117.
45. Ibid. p. 118.
46. Ibid. p. 118.
47. Ibid. p. 119-120.
48. Ibid. p. 121.
49. Ibid. p. 121.
50. Ibid. p. 121.
51. Ibid. p. 125.
52. Ibid. p. 126.
53. Garrison (1978), p. 167-204.
54. I am deeply grateful to Dr. Garrison for granting me permission to include an extensive summary of his argument. Of course, any errors of interpretation are solely my responsibility.
55. Ibid. p. 179.
56. Ibid. p. 179.
57. Ibid. p. 174.
58. Ibid. p. 172-173.
59. Ibid. p. 171.
60. Ibid. p. 173.

61. Ibid. p. 175.
62. Ibid. p. 175.
63. Ibid. p. 176.
64. Ibid. p. 177.
65. Ibid. p. 176.
66. Ibid. p. 184.
67. Ibid. p. 184.
68. Ibid. p. 185.
69. Ibid. p. 185.
70. Ibid. p. 186.
71. Ibid. p. 187-188.
72. Ibid. See Figure 8, p. 187.
73. Ibid. p. 188.
74. Ibid. p. 188.
75. Ibid. p. 190.
76. Ibid. p. 191.
77. Ibid. p. 191-192.
78. Snowdon, et al. (1994), p. 358-360.
79. Garrison (1978), p. 196.
80. Snowdon, et al. (1994), p. 358.
81. Ibid. p. 360.
82. Ibid. p. 360.
83. This equation may be found in Snowdon, et al. (1994), p. 369.
84. Dasgupta, Neha. "More than 25 Million People Apply for Indian Railway Vacancies." *Reuters*. Web. March 29, 2018. Accessed on April 5, 2018. https://www.reuters.com/article/us-india-unemployment-railways/more-than-25-million-people-apply-for-indian-railway-vacancies-idUSKBN1H524C

Chapter 15

THEORIES OF FINANCIAL MARKETS

Goals and Objectives:

In this chapter, we will do the following:

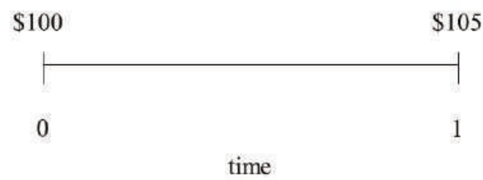
1. *Explain* how the rate of interest is defined and measured
2. *Explore* the relationship between the bond market and the loanable funds market
3. *Analyze* a neoclassical general equilibrium model of interest rate determination
4. *Incorporate* the stock market into the neoclassical theory of interest rate determination
5. *Investigate* an Austrian theory of interest rate determination
6. *Examine* a Marxian theory of interest rate determination

Prior to this chapter, our exploration of macroeconomic theory has been focused on theories of the business cycle. That is, we have concentrated mostly on factors that influence the overall amount of economic activity, the total production of commodities, and the amount of unemployment. In this chapter, we turn to theories of financial markets. The financial markets have an important role to play in market capitalist economies and if we are to gain a deeper understanding of macroeconomic policy in later chapters, we must first learn to think about how the financial markets work and how they interact with the rest of the economy. Once we have developed a more complete understanding of interest rates, bonds, and stocks, we will then be able to explore in detail the neoclassical, Austrian, and Marxian theories of interest rate determination.

The Definition and Measurement of the Rate of Interest

The **rate of interest**, or the interest rate, is simply an amount of money paid to a lender by a borrower for the use of money during a specific period, expressed as a percentage of the amount borrowed. For example, if a lender receives $5 in payment for the use of a $100 loan during a year, then the annual interest rate is 5% (= $5/$100). In this case, the $100 is referred to as the **principal** amount of the loan, and the $5 is the dollar amount of the interest. If the principal is returned at the end of one year, then that will be the end of the transaction. The lender will have received $105. This growth of the principal is captured in the simple diagram in Figure 15.1.

Figure 15.1: The Growth of Principal Over Time

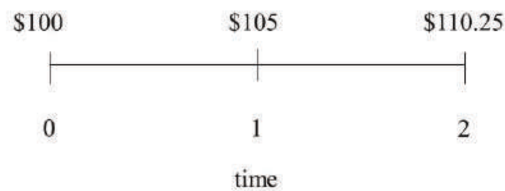


On the other hand, if the loan is renewed, then a situation of **compound interest** arises. That is, the lender leaves the $105 with the borrower at the end of the year, and then the lender expects to receive a 5% interest payment at the end of the second year calculated using the entire $105 loaned at the beginning of the second year. The calculation of the **future value** (FV) of the original loan amount of $100 at time t = 2 is as follows:

$$FV = 100(1 + 0.05)(1 + 0.05) = (100)(1.05)^2 = \$110.25$$

The growth of the principal using compound interest in this scenario is captured with the diagram in Figure 15.2.

Figure 15.2: The Case of Compound Interest for a 2-Year Loan

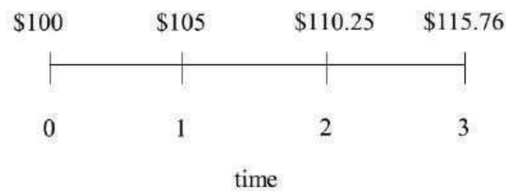


If the period of the loan is three years, then the future value will be even larger, again as a result of compound interest. The calculation of the future value in time t = 3 is as follows:

$FV = 100(1 + 0.05)(1 + 0.05)(1 + 0.05) = (100)(1.05)^3 = \$115.76$

The growth of the principal using compound interest in this scenario is captured with the diagram in Figure 15.3.

Figure 15.3: The Case of Compound Interest for a 3-Year Loan



In general, if the annual interest rate is i, the **present value** amount (or the initial loan amount) is PV, and the loan is made for n years, then the calculation of the future value is as follows:

$$FV = PV(1+i)^n$$

In the examples we just considered, a sum of money was loaned out and we explored how much it would be worth at the end of the loan period. That is, we considered the sum's future value. It is often the case, however, that we are confronted with different information. For example, we might know the future payment that is to be received in a known number of years. If we also know the interest rate, then we can calculate the present value of that sum by simply rearranging the future value formula as follows:

$$PV = \frac{FV}{(1+i)^n}$$

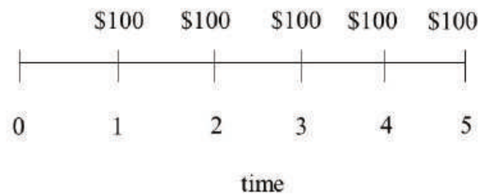
To use our earlier example, suppose that a lender knows she will receive $115.76 at the end of three years. If she knows that the annual interest rate is 5%, then she can arrive at the present value in the following way:
$$PV = \frac{FV}{(1+i)^n} = \frac{\$115.76}{(1+0.05)^3} = \$100$$

In other words, it is possible to equate future dollars with present dollars using the interest rate. The idea is rather intuitive. It means that future dollars are worth less than present dollars to a person. Wouldn't you rather have $100 today than in three years? Of course, you would. How large would the future sum need to be before you would consider it equivalent to the $100 today? According to the information reflected in the current rate of interest, the answer is $115.76.

When this method of determining the present value of a future sum is used, it is said that the future sum has been **discounted** to the present using the interest rate. The present value formula and the method of discounting are especially useful when we wish to know the present value of a specific financial asset. For example, a creditor (or lender) might purchase an asset, such as a bond. A **bond** is really just a financial contract between a lender and a borrower, much like an IOU. When a creditor purchases a newly issued bond, she hands over a sum of money to a borrower. The borrower agrees to repay the amount borrowed when the bond matures. In the case of **coupon bonds**, the borrower also agrees to make periodic payments of interest to the lender until the bond matures. If we know what the future payments will be, then we can discount each payment back to the present and then simply sum them up to determine the present value of the bond.

For example, suppose that a bond pays $100 in interest annually for the next five years. This situation is depicted in the diagram in Figure 15.4.

Figure 15.4: A 5-Year Bond with a Fixed Coupon Payment



If we know that the interest rate is 5% and we ignore the repayment of principal for simplicity, then we can calculate the present value of the bond as follows:

$$PV = \frac{\$100}{(1+0.05)} + \frac{\$100}{(1+0.05)^2} + \frac{\$100}{(1+0.05)^3} + \frac{\$100}{(1+0.05)^4} + \frac{\$100}{(1+0.05)^5} = \$432.95$$

For this calculation, the reader should notice that each future payment of $100 is being discounted back to the present before being summed up. Furthermore, each future payment is being discounted according to how many years will pass before it is received. That is, the final payment in year 5 is discounted the most (5 times). The payment at the end of year 1 is discounted the least (once). Finally, the reader should notice that the straightforward sum of the future payments is $500, but the present value is only $432.95. The reason, of course, is that the future payments are not worth $500 today due to the existence of discounting.

The important point to notice is that we can determine the present value of a bond or any financial asset if we know the future payments associated with the asset, the rate of interest, and the **term to maturity** (i.e., the number of years to maturity). Imagine that an investor is considering the purchase of several different assets, each with a different number of years to maturity and a different periodic payment. The only rational way to compare these different assets is to use the current interest rate to discount the future payments associated with each asset back to the present. Once the values of the different assets are determined for the present period, their values can be easily compared.

In these examples, we have been assuming that the interest rate is a known quantity. It is possible that an investor might be considering the purchase of a bond, but he only knows the amount of the periodic interest payments, the term to maturity, and the price of the bond. For example, suppose that the periodic payment is $100, the term is 5 years, and the price of the bond is $400. The investor would like to know the interest rate associated with the bond, which is also referred to as the **yield to maturity** (YTM). To calculate the yield, it is only necessary to determine the interest rate that will equate the current price of the bond with the present value of its future payments as follows:

$$400 = \frac{\$100}{(1+i)} + \frac{\$100}{(1+i)^2} + \frac{\$100}{(1+i)^3} + \frac{\$100}{(1+i)^4} + \frac{\$100}{(1+i)^5}$$

This calculation is difficult without a financial calculator. It can be obtained through a trial and error method. The solution is approximately a 7.93% rate of interest. Using this method, it is possible to measure the rate of interest that applies to a specific financial asset.

The reader should also notice that because the price of the bond is below $432.95, the yield on the bond is above 5%. That is, when an investor pays a lower price for the bond, with the periodic interest payments fixed, the yield is necessarily higher. This result is consistent with the widely reported relationship between interest rates and bond prices that one hears in the financial news: *Interest rates and bond prices are always inversely related.*

Economists of all persuasions tend to refer to the rate of interest as though it is a single entity. In reality, many different interest rates exist in market capitalist economies. Each corresponds to a different loan or asset. Interest rates exist for 12-month certificates of deposit, online savings accounts, 15-year mortgage loans, 30-year mortgage loans, 10-year Treasury bonds, 2-year auto loans, and so on. The reason that economists often refer to a single interest rate is that the many different interest rates that exist tend to move together. Economists do have theories as to how and why interest rates differ from one another, but frequently they are interested in explaining the overall movement of interest rates instead. In the latter case, they refer simply to "the interest rate."[1]

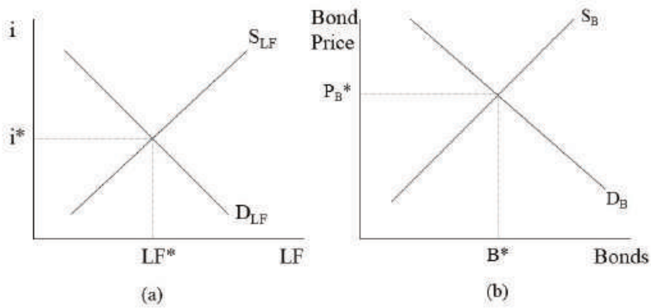
We are now in possession of a clear definition and method of measurement of the rate of interest. We have also learned how to calculate the future value of a present sum and the present value of a future sum. These tools will be very useful as we consider linkages between different financial markets in the next several sections of this chapter.

The Relationship between the Bond Market and the Loanable Funds Market

We are now in a position to identify the linkage between two key financial markets that exist in capitalist economies: the bond market and the loanable funds market. Although we discuss these markets as though they are two separate markets, each can be thought of as the mirror reflection of the other and so they are really one and the same.

The **loanable funds market** is the market for financial loans. Like any market, it has both a supply side and a demand side. On the supply side of the market are the lenders who possess loanable funds that they wish to lend to borrowers for a period of time in exchange for interest. The supply curve of loanable funds is upward sloping as shown in Figure 15.5 (a). That is, as the rate of interest increases, the quantity of loanable funds that lenders are willing and able to provide increases. The rate of interest may be thought of as the price paid for the use of these funds, and so the supply curve is upward sloping, as it is in most product markets. The demand curve for loanable funds is downward sloping, which is also shown in Figure 15.5 (a). It slopes downward because as the interest rate falls, more investment projects become profitable for businesses, and so businesses are willing and able to borrow more loanable funds.

Figure 15.5: The Linkage Between the Loanable Funds Market and the Bond Market



It should be clear that an **equilibrium interest rate** exists in this market. That is, the interest rate that occurs at the intersection of the two curves represents the interest rate that clears the market for loanable funds of both surpluses and shortages. If the interest rate is above the equilibrium interest rate, i*, then an excess supply or a surplus of loanable funds exists. In that case, more suppliers wish to lend than borrowers wish to borrow due to the high interest rate. As a result, competition among lenders will drive the interest rate down towards the equilibrium level. Alternatively, if the interest rate is below the equilibrium interest rate, i*, then an excess demand or a shortage of loanable funds exists. In that case, more borrowers wish to borrow than lenders wish to lend at the low interest rate. As a result, competition for the limited funds available will drive the interest rate up towards the equilibrium level. It should also be noted that an **equilibrium quantity of loanable funds**, LF*, is determined as well when the market is in equilibrium.

The second market, depicted in Figure 15.5 (b), is the **bond market**, which is the market for bonds. As before, when the price of bonds rises, the quantity of bonds that bondholders are willing and able to sell increases. Hence, the supply curve of bonds is upward sloping. When the price of bonds falls, on the other hand, the quantity of bonds that buyers are willing and able to purchase increases. As a result, the demand curve for bonds is downward sloping. As usual, the intersection of the two curves implies an **equilibrium bond price**, P_B*, and **an equilibrium quantity of bonds**, B*. The bond price adjusts to clear the market of surpluses and shortages, which establishes the equilibrium price and quantity exchanged of bonds.

It is interesting to note that the suppliers in the one market are the demanders in the other market. Similarly, the demanders in the one market are the suppliers in the other market. For example, if I wish to borrow funds in the loanable funds market, then I am on the demand side of that market. At the same time, I can only obtain loanable funds by selling bonds in this scenario, and so I am on the supply side in the bond market. Similarly, if I wish to lend funds in the loanable funds market, then I am on the supply side of that market. At the same time, I can only lend loanable funds by buying bonds in this scenario, and so I am on the demand side in the bond market. To summarize:

1. Demanders of loanable funds = Suppliers of bonds ⇒ Borrowers
2. Suppliers of loanable funds = Demanders of bonds ⇒ Lenders

Because of this logical connection between the two markets, we regard the loanable funds market as the mirror reflection of the bond market. It follows that if one of the markets is in equilibrium, the other market must also be in equilibrium. A natural question to ask then is whether a definite relationship exists between the equilibrium bond price and the equilibrium interest rate.

The answer to this question is that the equilibrium bond price will equal the present value of the bond calculated using the equilibrium interest rate. For example, suppose that we know the term to maturity for a bond to be 5 years, the initial loan amount to be A, and the equilibrium interest rate to be i*. Using this information, it is possible to calculate the present value of the bond when the loanable funds market is in equilibrium. It is this present value calculation that also yields the equilibrium price of the bond, P_B^*, as shown below:

$$P_B^* = PV = \frac{A}{(1+i^*)} + \frac{A}{(1+i^*)^2} + \frac{A}{(1+i^*)^3} + \frac{A}{(1+i^*)^4} + \frac{A}{(1+i^*)^5}$$

It might not be obvious why the price of the bond has an inherent tendency to move towards this level, P_B^*. The argument must be made, however, if the claim is to be defended that this bond price is an equilibrium one. To understand why this bond price represents the equilibrium bond price, consider what will happen if the price of the bond is different from this level. For example, suppose the price of the bond is above the present value as calculated here. In that case, investors will not want to purchase the bond. Who would pay more for a bond than its present value? The demand for the bond will fall, which places downward pressure on the bond price. Similarly, bondholders will be eager to sell the bond. Who will want to hold a bond when it can be sold at a price that exceeds its present value? The

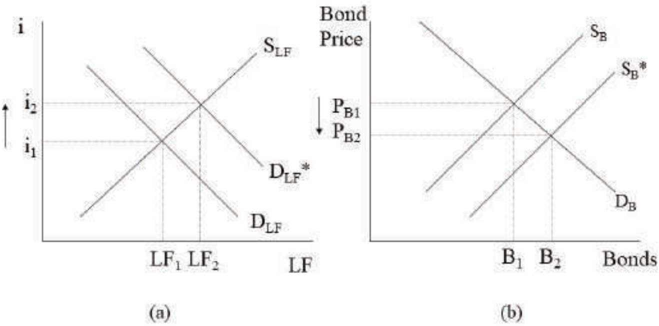
increase in supply will also put downward pressure on the bond price. Hence, both factors push the price of the bond down towards the present value of the bond and towards P_B^*. To summarize:
If $P_B^* > PV$ then demand falls and supply rises until $P_B^*=PV$.

Alternatively, consider what will happen if the price of the bond is below the present value as calculated here. In that case, investors will want to purchase the bond. Who would not want to purchase a bond when its price is less than its present value? It's a bargain. The demand for the bond will rise, putting upward pressure on the price of the bond. Similarly, bondholders will not want to sell the bond. Who would sell a bond when its price is below what it is worth in today's terms? The reduction in supply will put upward pressure on the bond price. Hence, both factors push the price of the bond up towards the present value of the bond and towards P_B^*. To summarize:
If $P_B^* < PV$ then demand rises and supply falls until $P_B^*=PV$.

Overall, we see that the equilibrium interest rate and the equilibrium bond price are related in a precise manner as reflected in the present value formula.

Finally, it is helpful to consider what will happen when equilibrium is disrupted in these markets. Suppose that the demand for loanable funds rises as a result of businesses becoming more optimistic about the health of the economy. A rightward shift of the demand curve for loanable funds will cause a rise in the equilibrium interest rate as shown in Figure 15.6 (a).

Figure 15.6: Simultaneous Equilibrium in the Bond Market and Loanable Funds Market



At the same time, because the demanders of loanable funds are also the suppliers of bonds, the supply curve of bonds will shift to the right in the bond market as shown in Figure 15.6 (b). The result will be a reduction in the equilibrium bond price. It is worth noting that this result of an inverse relationship between the interest rate and the bond price is consistent with the same conclusion drawn earlier from the present value formula.

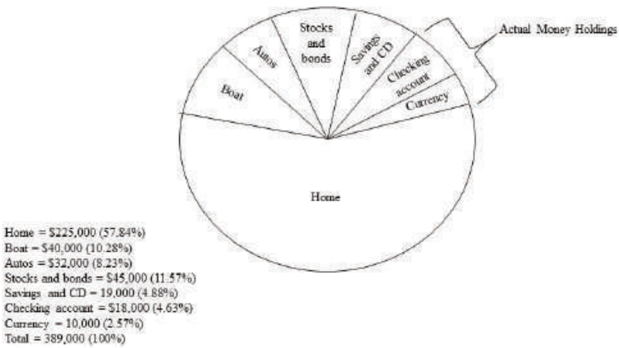
The Market for Money

In addition to the bond and loanable funds markets, mainstream economists typically also discuss a third financial market that is referred to as the **money market**. This concept can easily become the source of great confusion because of the different meanings of this phrase.[2] In the financial services industry, the money market refers to the market for short term securities. That is, assets with terms to maturity of less than one year are referred to as **money market instruments**. Three-month Treasury bills, six-month certificates of deposit, and commercial paper are all examples of assets that involve repayment of principal and interest to a lender in a period of less than one year. Investors purchase these assets to earn interest income using their short-term savings.

It is crucial to understand that the money market we will be discussing in this section is not the money market to which professionals in the financial services industry refer. Indeed, the money market to which financial services professionals refer is much closer to the bond market that we have been discussing, although many bonds have terms to maturity of much longer than one year. Instead, when mainstream economists refer to the money market, they have in mind a theoretical construct that derives from the work of John Maynard Keynes's 1936 book *The General Theory of Employment, Interest, and Money* that we discussed in Chapter 13. Because this "market" is rather unusual, we will need to devote some space to this notion. Once the money market is completely understood, we can complete the picture of the neoclassical theory of interest rate determination.

Consider for a moment, the finances of a single household. The household has accumulated a large collection of assets, which includes a home, two automobiles, a savings account, some stocks, some bonds, a certificate of deposit, a checking account, some currency (paper money and coins), and a boat. At any one point in time, the total value of the household's assets is divided into these different components, and it is possible to identify the dollar value that corresponds to each component, as shown in Figure 15.7.

Figure 15.7: The Division of a Household's Assets into Component Parts



Of all these different assets, neoclassical economists only consider the checking account and the currency to constitute money. **Money**, according to the neoclassical perspective, refers to anything that can be readily used for transactions. Only the most liquid assets qualify. **Liquidity** refers to the ease of conversion of an asset into currency. Currency is obviously the most liquid asset. Checking accounts are also extremely liquid. The funds are payable on demand and checks and debit cards can be easily used to engage in transactions. Savings deposits and certificates of deposit are less liquid because check writing is not possible and withdrawal restrictions apply to each. Stocks and bonds must be sold, which requires payment of a brokerage fee so they are not as liquid as currency or checkable deposits. The least liquid assets for the household are the home, the automobiles, and the boat. These must be sold, which takes time and is costly. Because currency and checkable deposits are the most liquid assets, neoclassical economists typically only consider them to be money.

Given this definition of money, the household's **actual money holdings** consist of the sum of its currency holdings and its checkable deposits. On the other hand, this household's **demand for money** refers to its desired money holdings, given its wealth. Of course, it is possible that the household's desired holdings and its actual holdings do not agree. If the household wishes to hold the same amount of money that it is holding in Figure 15.7, however, then this amount constitutes the household's demand for money.

It is important not to be confused by the concept of money demand. One might think that the demand for money should always be infinite because everyone always wants more of every good and asset, including money, according to the neoclassical way of thinking. This conclusion would be incorrect, however, because money demand only refers to desired money holdings *given the assets of the household*. That is, how much of the household's assets does it wish to hold in the form of money? It may wish to hold a lot of its assets in the form of money or only a little depending on the benefits it perceives to flow from the holding of money.

What are the benefits that flow from the holding of money? Why would a household hold any money? The classical economists considered this question and provided a helpful, albeit somewhat obvious answer. People hold money for the purpose of engaging in transactions. This **transactions demand for money** forms one part of the household's money demand. That is, the household requires money if it is to pay for goods and services. A household cannot survive for very long in a market capitalist society if it refuses to use money. Bills must be paid and groceries must be purchased. Clearly, the transactions demand for money seems to be an important piece of the puzzle.

Why else might a household decide to hold some of its assets in the form of money? Keynes offered an additional reason why a household might choose to hold money. Even if the household is not interested in using the money for planned transactions, it might wish to hold some money strictly for precautionary reasons. That is, the fear of unplanned medical expenses might lead a household to maintain an emergency savings fund, just in case. Fear of an unanticipated job loss might be another reason to hold wealth in the form of money. Neoclassical economists refer to this type of money demand as the **precautionary demand for money**. This factor might also be contributing to the money demand represented in Figure 15.7.

If we wish to draw the demand curves representing the transactions demand (D_T) and the precautionary demand (D_P) for money, we can do so as in Figure 15.8.[3]

Figure 15.8: The Transactions Demand and the Precautionary Demand for Money

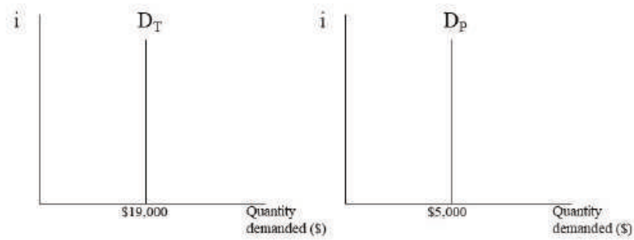


Figure 15.8 suggests that the sum of these two components of money demand (D_T+D_P) will be represented as a vertical line when the interest rate is placed on the vertical axis. That is, neither of these components of money demand depends on the interest rate. As the interest rate rises, the quantity of money demanded remains the same because households still wish to engage in the same volume of transactions and maintain the same precautionary balances in case of emergencies. When money income changes, however, the transactions demand and precautionary demand increase, and so these curves shift to the right. That is, households will want to purchase more goods and services, and they will prefer to hold more money in case of emergencies. A reduction in money incomes would lead to leftward shifts for similar reasons.

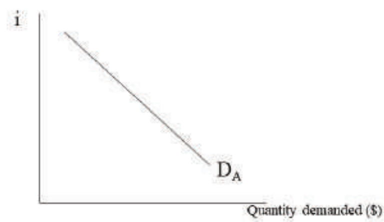
Finally, Keynes identified a third motive for holding money. He argued that a household might hold money purely for speculative reasons. That is, a household might wish to hold money so it might be used to purchase bonds if interest rates unexpectedly rise to higher-than-normal levels. Due to the inverse relationship between bond prices and interest rates, the drop in bond prices will make them attractive investments. Once interest rates fall and bond prices rise to their original, normal levels, the household will enjoy a capital gain. A **capital gain** is the difference between the selling price and the purchase price of an asset. This **speculative demand for money** offers a third way to understand the demand for money represented in Figure 15.7.

Although Keynes wrote about the speculative demand for money, neoclassical economists often think in terms of the **asset demand for money**, which is like the speculative demand for money.[4] If money is thought about as an asset, then it can be compared to the other assets the household owns. Initially, it might appear that money is a terrible asset. A home is a great long-term investment because home prices frequently increase over time at a rate greater than most prices. As a result, a homeowner might experience a capital gain when she finally sells the house, and in the meantime, the members of the household have been able to enjoy the benefits of living in the home. Stocks and bonds may also be sold to realize capital gains, and they pay dividends and interest, respectively, while they are owned. Even savings accounts and CDs pay interest to their owners. Money, however, generally pays no interest. Currency in your wallet and checkable deposits do not pay interest.

What does money have to recommend it then as an asset? The answer: its liquidity. Assets are not only evaluated on the basis of their expected return. Their degree of liquidity is also an important characteristic that is important to investors. Because money is the most liquid asset available, it is typically held as an asset. Remember, if this component of money demand exists, then a part of the total money held by the household is not for the purpose of making transactions or for precautionary reasons. It is held simply because money is one asset among several that the household considers to be worth holding. The asset demand for money depends on a household's money (or nominal income), just like the transactions demand and the precautionary demand. As a household's money income rises, its wealth increases, and it will choose to hold more of all assets, including money.

The asset demand for money also depends, however, on another key factor: the rate of interest. As the rate of interest rises, interest-bearing assets become relatively more desirable. That is, the opportunity cost of holding money as an asset increases. Households, therefore, wish to hold less money, and the quantity of money demanded declines. Alternatively, as the rate of interest declines, the quantity of money demanded increases because the opportunity cost of holding it falls. That is, interest-bearing assets become relatively less attractive, and the liquidity characteristic of money makes it seem like a relatively more attractive asset. If we place the rate of interest on the vertical axis and the quantity of money demanded on the horizontal axis, then the curve representing the asset demand for money (D_A) is downward sloping, indicating an inverse relationship between the quantity of money demanded and the interest rate. This situation is depicted in Figure 15.9.

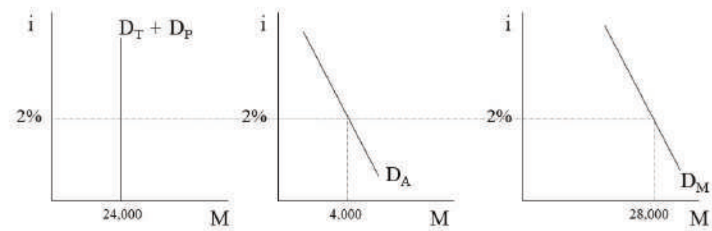
Figure 15.9: The Asset Demand for Money



A rise in money income will shift the money demand curve to the right because of the positive affect on wealth and the demand for all assets. Its downward slope, however, is entirely the result of the negative relationship between money demand and the interest rate.

We are now able to represent the entire money demand curve that captures all three components of money demand. Using the same approach of *horizontal summation* that we used in Chapter 3, it is possible to aggregate the transactions demand and precautionary demand curve for money and the asset demand curve for money to obtain the total money demand curve (D_M), as shown in Figure 15.10.

Figure 15.10: The Total Demand for Money

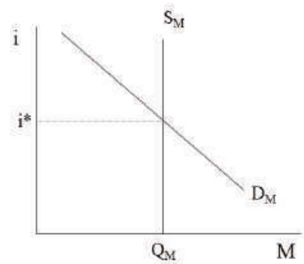


The total money demand curve will be downward sloping as a result of the downward slope of the asset money demand curve. Because it represents the sum of all three curves, it lies further to the right than any one of the curves taken individually. As before, a change in money income will shift the money demand curve in the direction of the change.

Up until this point, we have been discussing the demand for money originating with a single household. Of course, each household will have its own money demand curve, reflecting its desire to hold money for transactions, for precautionary reasons, and as an asset. If we aggregate all these individual households' demands and all the firms' demands for money (after all, businesses will desire to hold money as well for various reasons), then we obtain the aggregate money demand curve for the entire economy. It should also be downward sloping for the reasons described in this section. Using the aggregate money demand curve, we now have a completely developed notion of one side of the market that neoclassical economists call the money market.

The demand side of the money market refers to desired money holdings. Actual money holdings, however, are reflected in the supply side of the money market. The supply of money is collectively determined by the nation's central bank, the commercial banks, and the depositors. Among these three, the central bank has the greatest influence over the money supply. In the United States, the Federal Reserve (referred to as "The Fed") serves as the central bank. It has the power to determine the quantity of checkable deposits and currency in circulation. In other words, the Fed determines the quantity of money supplied. At this stage, we assume that the money supply is exogenously determined. That is, the money supply is determined by central bankers, and it is independent of the interest rate. Therefore, the money supply curve (S_M) is perfectly vertical as shown in Figure 15.11.

Figure 15.11: The Money Market



If the Fed increases the money supply, then the money supply curve shifts to the right. If the Fed reduces the money supply, then the money supply curve shifts to the left.

It is also worth noting that the supply and demand curves intersect at a specific interest rate (i*) in the money market. At this interest rate, the quantity of money that households and firms actually hold equals the quantity that households and firms desire to hold. That is, actual money holdings equal desired money holdings. In this situation, households and firms have no reason to modify their behavior, and so this interest rate represents the equilibrium rate of interest. What is not clear at this stage is how equilibrium is achieved in this market. We provide the answer to this question in the next section.

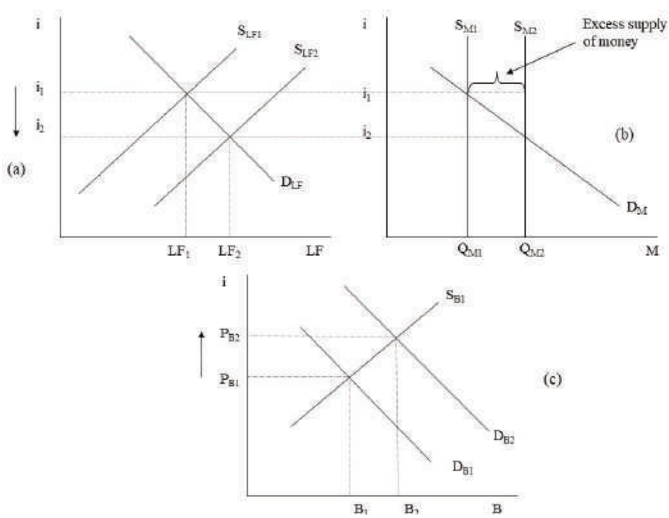
A Neoclassical General Equilibrium Model of Interest Rate Determination

In this section, we will consider a neoclassical general equilibrium model of interest rate determination. A **general equilibrium model** is a model that demonstrates how multiple markets simultaneously arrive at an equilibrium outcome. In earlier chapters, our focus was on partial equilibrium models. A **partial equilibrium model** demonstrates how a single market reaches an equilibrium outcome. Because partial equilibrium models are very easy to explain, neoclassical economists are fond of using such models at the introductory level. When discussing financial markets, however, it is helpful to use a general equilibrium framework that links together the loanable funds market, the bond market, and the money market. It can be shown that a change in any one market leads to the clearing of the other markets.

The first economist to rigorously develop a mathematical model of general equilibrium was the French economist Leon Walras. In the 1870s, Walras was one of the three economists to independently emphasize marginal changes as central to rational economic decision making. William Stanley Jevons in Britain and Carl Menger in Austria were the other two economists to participate in what would later be dubbed the **marginalist revolution**. The new approach to economic analysis was considered significant enough that later economists would regard this change as an event separating old-fashioned classical economics from modern neoclassical economics. Walras's work was unique, however, in that he also developed a theory of general economic equilibrium.

An important part of Walrasian general equilibrium theory is something called Walras's Law. **Walras's Law** states that if n markets exist and n – 1 markets are in equilibrium, then the nth market must also be in equilibrium. Technically, one market cannot be out of equilibrium while the other markets remain in equilibrium. To illustrate the concept of simultaneous equilibrium in multiple markets, however, we will consider the simultaneous adjustments that occur to restore general equilibrium when one market is thrown out of equilibrium in our simple model of three financial markets. To explore how these markets adjust, let's consider what happens when all three markets begin in equilibrium, but then an exogenous shock causes equilibrium in the money market to be disrupted. For example, suppose that the Fed increases the money supply. In this case, the money supply curve shifts to the right, as shown in Figure 15.12 (b).

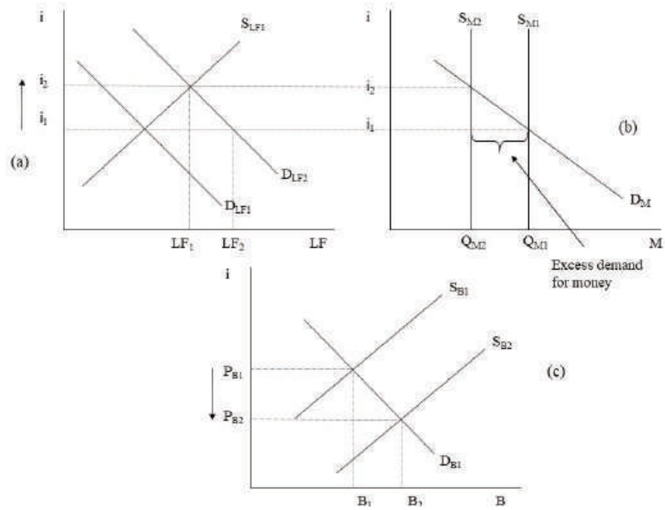
Figure 15.12: An Increase in the Money Supply



As the reader can see, the rightward shift of the money supply curve creates a surplus of money in the money market at the original interest rate, i_1. That is, households and firms are now holding more money than they wish to hold at the current interest rate. As a result, they will use the surplus funds to buy bonds, thereby increasing the demand for bonds from D_{B1} to D_{B2} as shown in Figure 15.12 (c). The increased demand for bonds drives up the price of bonds towards its new equilibrium level. At the same time, the higher demand for bonds is equivalent to an increase in the supply of loanable funds, and so the supply of loans shifts from S_{LF1} to S_{LF2} as shown in Figure 15.12 (a). The interest rate falls towards its new equilibrium level of i_2. As the interest rate falls in the loanable funds market, it also falls in the money market, and so an increase in the quantity demanded of money occurs, represented as a movement along the money demand curve. This movement continues until the money market is also in equilibrium. The result is simultaneous equilibrium in all three markets.

Now let's return to the initial situation of equilibrium in all three markets, but this time let's suppose that the Fed reduces the money supply, shifting the money supply curve to the left, as shown in Figure 15.13 (b).

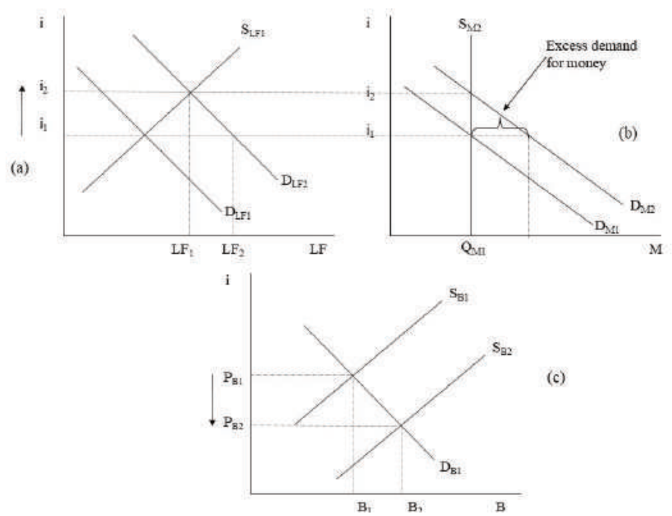
Figure 15.13: A Decrease in the Money Supply



This time the leftward shift of the money supply curve creates a shortage of money in the money market at the original interest rate, i_1. That is, households and firms are now holding less money than they wish to hold at the current interest rate. As a result, they will attempt to acquire funds by selling bonds, thereby increasing the supply of bonds from S_{B1} to S_{B2} as shown in Figure 15.13 (c). The increased supply of bonds drives down the price of bonds towards its new equilibrium level. At the same time, the higher supply of bonds is equivalent to an increase in the demand for loanable funds, and so the demand for loans shifts from D_{LF1} to D_{LF2} as shown in Figure 15.13 (a). The interest rate rises towards its new equilibrium level of i_2. As the interest rate rises in the loanable funds market, it also rises in the money market and so a decrease in the quantity demanded of money occurs, represented as a movement along the money demand curve. This movement continues until the money market is also in equilibrium. Again, the result is simultaneous equilibrium in all three markets.

Once again, let's return to the initial situation of equilibrium in all three markets, but this time let's suppose that money incomes increase so that the demand for money rises. In this case, it is a rightward shift of the money demand curve that occurs as shown in Figure 15.14 (b).

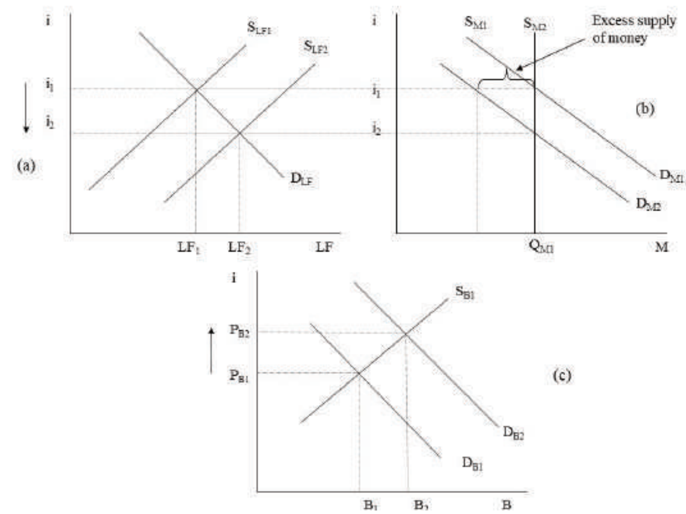
Figure 15.14: An Increase in the Demand for Money



In this case, a rightward shift of the money demand curve creates a shortage of money in the money market at the original interest rate, i_1. That is, households and firms are now holding less money than they wish to hold at the current interest rate. As a result, they will attempt to acquire funds by selling bonds, thereby increasing the supply of bonds from S_{B1} to S_{B2} as shown in Figure 15.14 (c). The increased supply of bonds drives down the price of bonds towards its new equilibrium level. At the same time, the higher supply of bonds is equivalent to an increase in the demand for loanable funds, and so the demand for loans shifts from D_{LF1} to D_{LF2} as shown in Figure 15.14 (a). The interest rate rises towards its new equilibrium level of i_2. As the interest rate rises in the loanable funds market, it also rises in the money market, and so a decrease in the quantity demanded of money occurs, represented as a movement along the money demand curve. This movement continues until the money market is also in equilibrium. Again, the result is simultaneous equilibrium in all three markets.

Finally, let's return to the initial situation of equilibrium in all three markets, but this time let's suppose that money incomes decrease so that the demand for money falls. In this case, it is a leftward shift of the money demand curve that occurs as shown in Figure 15.15 (b).

Figure 15.15: A Decrease in the Demand for Money



In this case, the leftward shift of the money demand curve creates a surplus of money in the money market at the original interest rate, i_1. That is, households and firms are now holding more money than they wish to hold at the current interest rate. As a result, they will use their surplus funds to purchase bonds, thereby increasing the demand for bonds from D_{B1} to D_{B2} as shown in Figure 15.15 (c). The increased demand for bonds drives up the price of bonds towards its new equilibrium level. At the same time, the higher demand for bonds is equivalent to an increase in the supply of loanable funds, and so the supply of loans shifts from S_{LF1} to S_{LF2} as shown in Figure 15.15 (a). The interest rate falls towards its new equilibrium level of i_2. As the interest rate falls in the loanable funds market, it also falls in the money market, and so an increase in the quantity demanded of money occurs, represented as a movement along the money demand curve. This movement continues until the money market is also in equilibrium. As before, the result is simultaneous equilibrium in all three markets.

All the cases considered clearly show that any disruption in the money market will lead to adjustments in the bond and loanable funds markets to restore equilibrium in all three markets. The analyses are also consistent with the general observation that bond prices and yields are inversely related, and with the money market in equilibrium, all firms and households are holding precisely the amount of money that they desire to hold, given their total assets.

Incorporating the Stock Market into the Analysis

At this stage, it should be clear how interest rates are explained within the neoclassical framework. Households and firms compare their actual money holdings with their desired money holdings and then buy or sell bonds when any discrepancies exist. Eventually, the rate of interest adjusts, which brings desired money holdings into line with actual money holdings. No surpluses or shortages exist in any of the financial markets, and the situation will persist unless an external shock disrupts the general equilibrium.

One financial market that has not been incorporated into the analysis thus far is the stock market. Just like with bonds, corporations issue shares of stock in order to raise funds. Unlike bonds, however, s**tocks** are shares of ownership in the corporations that issued them. For example, Microsoft sells stock to the public. When a corporation sells stock to the public for the first time, it typically hires an investment bank, like Goldman Sachs or Morgan Stanley, to **underwrite** the stock issue. That is, it guarantees Microsoft a price per share and then sells it to the public in the **primary market**, pocketing a **promoter's profit** in the process.

The buyer of Microsoft stock becomes a part-owner of the corporation and thus has a partial claim to the net income and assets of the firm. If the firm fails, the stockholder loses her investment. If the firm makes profits, then the stockholder may receive profit distributions in the form of **dividends**. The stockholder might also decide to sell the stock in the **secondary market**, such as the New York Stock Exchange. If the price of the stock has increased since the time it was purchased, then the stockholder will enjoy a capital gain upon selling it.

One of the benefits of stock ownership is the right to participate in shareholders' meetings and to vote in elections that will decide the corporation's board of directors. The possibility of large dividends and capital gains also makes stock ownership attractive, but losses may be considerable as well, and so stocks are generally rather risky. Another downside to stock ownership is that bondholders have a prior claim to the assets of the firm. If the firm fails, then the stockholders will be the last individuals to receive a share of the failing firm's remaining assets.

Because many different companies issue their own stocks, when we refer to the stock market, we are referring to the market for many different financial assets. Just like we simplified our analysis in the previous section by referring to the bond market as the market for a single type of bond (even though many different types of bonds exist), we will discuss the stock market as the market for a single stock. Because bond prices tend to rise and fall together and stock prices tend to rise and fall together, many economists are comfortable developing theories without letting these differences stand in the way.

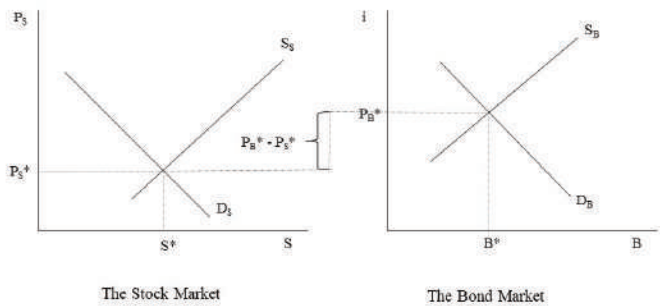
For simplicity, let's suppose that a share of stock may be purchased at a price P_S. To further simplify, let's assume that annual dividend payments are expected for the next five years but that the firm will cease to exist at the end of that time period. If the annual expected dividend payments are D for the next 5 years, then we can write an equation that allows us to determine the discount factor (d) for the stock in the following way:[5]

$$P_S = \frac{D}{(1+d^*)} + \frac{D}{(1+d^*)^2} + \frac{D}{(1+d^*)^3} + \frac{D}{(1+d^*)^4} + \frac{D}{(1+d^*)^5}$$

A strong similarity exists between this formula and the one used to determine the yield on a five-year bond. The key difference here is that the dividend payments are not guaranteed but only expected. That is, a bondholder knows the interest payments that she will receive (if the corporation does not **default** on the payments, which means it fails to pay the bondholder). For the stockholder, the payments are much less certain. They may be higher or lower than D. They may be suspended if the Board of Directors decides that reinvestment of the firm's profits is a superior move. Any number of factors might disrupt the payment of these expected dividends. As a result, stocks are inherently riskier than bonds and stockholders demand a premium to compensate them for this additional risk. The discount factor, is therefore, higher than the interest rate that we used to discount the interest payments associated with a bond. The discount factor applied to stocks (d*), in other words, will be the interest rate on bonds (i*) plus some additional amount to compensate for risk (i.e., a risk premium). That is, $d^* > i^*$. If the annual expected dividend payments (D) equal the fixed interest payments (A) on a bond issued by the same company (i.e., D = A), then the price of the stock will be lower than the price of the bond. That is, $P_S < P_B$. In other words, other factors the same, a stock should have a lower price than a bond due to its greater risk. The expected yield for the stock would then be higher than the yield on the bond. For this reason, stocks are generally viewed as *potentially* more lucrative than bonds.

To represent this situation graphically, we may consider the supply and demand for bonds and the supply and demand for stocks. Because stocks are riskier, having uncertain dividend payments, the demand for stocks will be lower than the demand for bonds. The equilibrium price of stocks (P_S*) will, therefore, be lower than the equilibrium price of bonds (P_B*), as shown in Figure 15.16, assuming other factors are the same, such as the expected payments associated with each asset.

Figure 15.16: The Discrepancy Between Stock Prices and Bond Prices



One might expect this situation to be temporary. If stocks are cheaper, then bondholders might be expected to sell their bonds and buy stocks instead. If bondholders behave in this way, then bond prices will fall and stock prices will rise until they are equal. This result will not occur, however, because the price differences are driven by the lower demand for stocks due to their greater perceived risk. The equilibrium is, therefore, stable, and the price discrepancy ($P_B^* - P_S^*$) will persist if the perceived degree of riskiness of the stock does not change.[6]

Real Interest Rates versus Nominal Interest Rates

One additional aspect of the neoclassical theory of interest rates that deserves attention is the neoclassical distinction between nominal interest rates and real interest rates, which was introduced in Chapter 12. The **nominal interest rate** is the interest rate that is actually observed in the marketplace. As stated previously, many different interest rates are observed in the marketplace, and each interest rate is a nominal interest rate. Because these nominal interest rates tend to move together over time, economic theories frequently refer only to the nominal interest rate.

The **real interest rate**, on the other hand, refers to the interest rate corrected for inflation. For example, suppose that you lend $100 to someone at a 5% nominal rate of interest. In one year, your money will have grown to $105. If the prices of goods have increased in the meantime, however, then the purchasing power of your money may not have increased at all. The **purchasing power** of money refers to its *real* value, or the amount of actual goods it can purchase. The real interest rate then refers to the percentage increase in the purchasing power of a sum of money during the period of a loan.

To calculate the real interest rate, we assume that a sum of money (M) is loaned to an individual who is charged the nominal interest rate (i). We also assume that the general price level, as measured by the GDP Deflator or the Consumer Price Index (CPI), is denoted as P. The real value of M at the time the loan is made is M/P. For example, if M is equal to $10 and a specific type of apple priced at $0.50 each is the only good produced in the economy, then the real value of M is 20 apples (= $10/$0.50 per apple). Of course, when using a price index, the real value of the money will be stated in terms of constant base year dollars.

Now let's suppose that M/P is loaned to a borrower. The real interest rate (r) will be the percentage change in M/P over the course of the year. The reader might know that

when considering the percentage change in the ratio of two variables, it is possible to estimate the solution by subtracting the percentage change in the denominator from the percentage change in the numerator:
$r = \%\Delta\frac{M}{P} \approx \%\Delta M - \%\Delta P$

This shortcut method of calculating the real interest rate has a helpful interpretation. The percentage change in M is simply the nominal interest rate (i) that the borrower is charged. The percentage change in P is the rate of inflation (π). Therefore, the real interest rate may be calculated as follows:
$r = i - \pi$

That is, the real interest rate is simply the difference between the nominal interest rate and the inflation rate. Therefore, if $100 is loaned out for one year at a 5% nominal interest rate and the inflation rate for that year is also 5%, then the real interest rate is 0% (= 5% – 5%). That is, the rising price level completely wipes out the nominal increase in the sum of money, leaving the lender no better off and no worse off than before. Because of the tendency for inflation to wipe out the gains from lending, lenders will aim to charge a nominal interest rate that is high enough to cover the inflation rate and then allow for a positive real interest rate. Of course, the real interest rate will be determined competitively in the financial markets, as we have discussed.

The reader should note that it is possible for the real interest rate to be negative. Even if the nominal interest rate is 5%, the real interest rate will be -3% if the inflation rate is 8%. In that case, the lender will receive back a sum of money that is nominally larger than the amount originally loaned out (by 5%), but it will have 3% less purchasing power due to the high inflation rate.

An Austrian Theory of Interest Rate Determination

The neoclassical school of economic thought offers only one among several approaches to interest rate determination. The Austrian school of economics has also made important contributions to the theories of capital and interest. Eugene von Bohm-Bawerk developed a theory of capital in the late nineteenth century that served as a theory of interest rate determination. His theory was based on the physical productivity of capital whereas Ludwig von Mises in the early twentieth century focused more on the role of subjective preferences in the formation of the rate of interest. In this section, we will look at one way of representing Mises's theory of interest within the context of a pure exchange economy. Because this theory can be developed without any reference to the theory of production, it serves as a straightforward introduction to the Austrian theory of interest.[7]

In developing his theory of interest, Mises relied heavily upon the concept of time preference. As described in Chapter 14, t**ime preference** refers to the fact that consumers have specific preferences regarding the time pattern of consumption. A consumer with a **positive time preference** prefers to consume in the relatively near future as opposed to the relatively distant future. A consumer with a **negative time preference** prefers to consume in the relatively distant future as opposed to the relatively near future. Finally, a consumer with a **neutral time preference,** or a time preference of zero, is indifferent between consuming in the relatively near future and the relatively distant future.[8] For example, suppose that three consumers have specific preferences regarding how many apples to consume in each of five periods of time. Table 15.1 represents the time preferences of the three consumers.

Table 15.1: The Time Preferences of Three Consumers

Desired Consumption in Each Period (in Number of Apples)			
Period	Consumer 1 (Positive)	Consumer 2 (Negative)	Consumer 3 (Neutral)
1	8	1	3
2	6	3	3
3	4	4	3
4	2	5	3
5	1	7	3

Clearly, consumer 1 has a positive time preference because she would like to consume more apples in the current period and in the relatively near future and fewer applies in the relatively distant future. Consumer 2 has a negative time preference. His desired consumption increases in the relatively more distant future periods. Finally, consumer 3 is comfortable consuming the same number of apples in every period regardless of when the apples will be consumed.[9] Although we can represent all of these cases, most consumers possess positive time preferences. That is, consumers prefer to consume now rather than later, but because these preferences are entirely subjective, negative and neutral time preferences are theoretically possible.

To further develop the Austrian theory of interest, let's consider an example in which three consumers receive endowments of apples in each of seven periods with the first period representing the current period. Table 15.2 shows the time pattern of endowments for each of the three consumers over the course of the seven time periods.

Table 15.2: The Time Pattern of Endowments for Three Consumers

	Endowments Per Period (in apples)		
Period	Consumer 1	Consumer 2	Consumer 3
1	5	4	2
2	6	5	3
3	3	7	6
4	2	2	5
5	7	3	5
6	5	6	4
7	4	5	7
Total	32	32	32

Of course, each consumer might have a time preference that differs substantially from their time pattern of endowments as shown in Table 15.3, which allows us to compare the time pattern of endowments with desired consumption for each consumer.

Table 15.3: The Time Pattern of Endowments and Desired Consumption for Three Consumers

	Endowments Per Period			Desired Consumption Per Period		
Period	C1	C2	C3	C1	C2	C3
1	5	4	2	8	9	10
2	6	5	3	7	8	7
3	3	7	6	6	5	5
4	2	2	5	5	4	4
5	7	3	5	3	3	3
6	5	6	4	2	2	2
7	4	5	7	1	1	1
Total	32	32	32	32	32	32

Rational people will try to bring their actual consumption into line with their desired consumption. If we assume zero storage costs, then a person with a negative time preference will be able to solve this problem herself. She simply needs to save her apples until later periods, assuming that the endowments are large enough in the periods closer to the present to allow for this plan to be executed. Since most people possess positive time preferences (like the three consumers in Table 15.3), however, consumers will typically want to transfer future apples to the time periods that are closer to the present. The problem is that apples cannot be transferred from the future to the present. A problem of time asymmetry exists. **Time asymmetry** means that present goods can be transferred to the future, but future goods cannot be transferred to the past.[10] The problem of time asymmetry means that consumers must find another method of increasing present consumption at the expense of future consumption. The way that consumers achieve this goal is by selling paper claims to their future apples to other consumers in the present. No consumer will pay one apple for a claim to one future apple because the consumer will gain nothing. A consumer will only give one apple in exchange for a paper claim to a future apple if the consumer is promised something extra as well. This something extra is called interest. Therefore, interest exists as a direct result of the presence of positive time preferences and the problem of time asymmetry. The market for paper claims to future production (i.e., the bond market) is also a direct result of subjective preferences for future goods.

To determine desired net borrowing for one consumer in a specific period, we only need to calculate the difference between her desired level of consumption and her endowment.[11] That is:

Desired Net Borrowing = Desired Consumption per Period – Endowment per Period

Table 15.4 shows the amount of desired net borrowing per period for each consumer. This amount of net borrowing is required for each consumer to achieve her desired consumption per period.

Table 15.4: The Desired Net Borrowing Per Period for Three Consumers

	Endowments Per Period			Desired Consumption Per Period			Desired Net Borrowing Per Period		
Period	C1	C2	C3	C1	C2	C3	C1	C2	C3
1	5	4	2	8	9	10	3	5	8
2	6	5	3	7	8	7	1	3	4
3	3	7	6	6	5	5	3	-2	-1
4	2	2	5	5	4	4	3	2	-1
5	7	3	5	3	3	3	-4	0	-2
6	5	6	4	2	2	2	-3	-4	-2
7	4	5	7	1	1	1	-3	-4	-6
Total	32	32	32	32	32	32	0	0	0

Table 15.4 shows that each consumer wishes to borrow in the current period and in the periods that are closer to the present time (for the most part). They would like to lend in the later periods as reflected in the negative desired net borrowings in the later periods. The problem that each consumer faces, however, is the constraint that the total consumption per period across all three consumers cannot exceed the total endowment per period for all three consumers. Trades of apples for paper claims can occur in a given period, but future apples cannot be received in the present. Any apples obtained above one's endowment can only be the result

of another consumer consuming an amount below her endowment. The situation depicted in Table 15.4 is impossible because the total desired consumption in period 1 is 27 apples, for example, but the total endowment in period 1 is only 11 apples. Therefore, this desired allocation is impossible. Another way of stating the same result is that the total *actual* net borrowing per period can never exceed zero. Any net borrowing must be balanced by net lending. In Table 15.4 desired net borrowing per period is positive for all three consumers. It is not possible for all three consumers to borrow in a specific period, and so these plans cannot be satisfied, and this ideal situation cannot arise.

The consumers resolve these difficulties to the best of their abilities by creating a bond market and supplying and demanding paper claims to future apples. The supply and demand for these paper claims will determine an equilibrium rate of interest. Let's assume that this problem is resolved in all periods and that the interest rate is 100% in every period with all interest paid in the very last period. This assumption is a greatly simplifying assumption that allows us to represent one possible result. After all this haggling occurs, the actual consumption per period for each consumer will be determined as shown in Table 15.5.

Table 15.5: The Actual Consumption, Actual Net Borrowing, and Final Tally

	Actual Consumption Per Period (omitting net interest)			Actual Net Borrowing Per Period			Actual Consumption Plus Net Interest		
Period	C1	C2	C3	C1	C2	C3	C1	C2	C3
1	2	5	4	-3	1	2	2	5	4
2	7	0	7	1	-5	4	7	0	7
3	6	5	5	3	-2	-1	6	5	5
4	4	3	2	2	1	-3	4	3	2
5	5	5	5	-2	2	0	5	5	5
6	2	4	9	-3	-2	5	2	4	9
7	3	8	5	-1	3	-2	**9**	**12**	**-5**
Total	29	30	37	-3	-2	5	35	34	27

The actual consumption per period just represents a redistribution of the total endowment in each period across the three consumers. It is determined as the consumers decide whether to borrow or lend in each period. The actual net borrowing per period for each consumer can also be calculated as follows:

Actual Net Borrowing = Actual Consumption per Period – Endowment per Period

In this example, the total actual net borrowing per period is equal to zero in each period. Negative values for actual net borrowing indicate net lending and positive values indicate net borrowing. Table 15.2 shows that the aggregate endowment is 96 apples. In Table 15.5 the grand total for actual consumption is also 96 apples.[12] The final three columns of Table 15.5 also show how the receipt of net interest in the final period affects the final results. Because the interest and principal are paid in period 7, those numbers are indicated in bold. The bold numbers represent the only difference between actual consumption per period plus net interest and actual consumption per period omitting net interest, as shown in Table 15.5. Net interest is calculated as follows:

Net Interest = Principal and Interest Received – Principal and Interest Paid

For example, to calculate the net interest for consumer 1 in period 7, simply multiply 2 times 9 (=3+2+3+1) to obtain 18 apples. This amount is the principal and interest received. It is obtained by multiplying each apple loaned out by 2, shown as negative net borrowings for consumer 1. Remember the interest rate is assumed to be 100% and so the multiplication by 2 ensures that we account for principal and interest. Next determine the principle and interest paid by consumer 1. Multiply each apple borrowed by 2, shown as positive net borrowings for consumer 1. That is, multiply 2 times 6 (=1+3+2) to obtain 12 apples. Finally, using these results, add 3 apples of actual consumption (found in the column that omits net interest) to 18 apples (principal and interest received), and then subtract 12 apples (principal and interest paid). The result is 9 apples.

In the end, the total actual consumption plus net interest across all three consumers is equal to 96 apples. Hence, all apples in the initial endowment are reallocated across time to improve the situation of each consumer, although no consumer achieves their most desired allocation. It is also worth noting that consumers 1 and 2 experience negative net borrowings overall, which means they are net lenders. Consumer 3, on the other hand, is a net borrower with a positive total of actual net borrowings. What this means is that, once the net interest is paid, consumers 1 and 2 end up with a total number of apples in excess of their original endowments of 32 apples each. Consumer 3, on the other hand, ends up with 27 apples and thus consumes less than the total endowment.[13] The reason, of course, is that consumer 3 is a net debtor and must pay a fair amount in interest for consuming the most in the first three periods.

The model presented here takes certain liberties in extending Mises's theory to illustrate a number of key points. The main point should be clear, however, which is that interest arises inevitably as a result of the different subjective time preferences of consumers. Because of the problem of time asymmetry, it is not possible for consumers to transfer future goods to the present. Hence, they engage in trade with one another in the present period and create a market for paper claims to future goods. A positive rate of interest is inevitable because otherwise no one would have an incentive to hand over present goods to another person. The rate of interest that results is determined competitively in the marketplace through a process of haggling. This representation of the Austrian theory of interest does not show how this process leads to the formation of a market rate of interest, but it does demonstrate that such a rate of interest arises because consumers seek to fulfill their consumption plans over time.

A Marxian Theory of Interest Rate Determination

Marxian economists also have ideas about how interest rates are determined. Karl Marx wrote about the formation of the rate of interest in Volume 3 of *Capital,* which was published after Marx's death by his friend and collaborator Frederick Engels in 1894. In this section, we will examine one interpretation of Marx's interest rate theory.[14]

In Marxian economics, the rate of interest is related in a logical way to the rate of profit (p) and another rate that Theodore Lianos calls the rate of profit of enterprise (r_e). To demonstrate the relationship between these three rates, it is necessary to begin with an identity. That is, aggregate profit (P) in the economy is identically equal to aggregate interest (I) plus aggregate **profit of enterprise** (R_e) as follows:

$$P = I + R_e$$

In other words, capitalist enterprises possess a total amount of profit, part of which is used to pay interest for the use of borrowed money capital. The other part is profit kept by the enterprise for its own internal use or for distribution to shareholders. The next step is to divide both sides by the aggregate capital, which consists of variable capital (V) and constant capital (C) as follows:

$$\frac{P}{C+V} = \frac{I}{C+V} + \frac{R_e}{C+V}$$

Next we multiply each expression on the right hand side of the equation by ratios that are equal to 1, thus maintaining the equality. That is:

$$\frac{P}{C+V} = \frac{I}{C+V} \cdot \frac{A}{A} + \frac{R_e}{C+V} \cdot \frac{C+V-A}{C+V-A}$$

A in this equation refers to the total borrowed money capital. Therefore, C+V–A refers to the total non-borrowed money capital. Rearranging the terms a bit yields the following result:

$$\frac{P}{C+V} = \frac{I}{A} \cdot \frac{A}{C+V} + \frac{R_e}{C+V-A} \cdot \frac{C+V-A}{C+V}$$

In this case, A/(C+V) refers to the fraction of the total capital that is borrowed. Similarly, (C+V-A)/(C+V) refers to the fraction of the total capital that is not borrowed. If we set k = A/(C+V), then the equation may be written as follows:

$$\frac{P}{C+V} = \frac{I}{A} \cdot k + \frac{R_e}{C+V-A} \cdot (1-k)$$

As is explained in Chapter 4, P/(C+V) represents the **rate of profit** for the economy as a whole. In Chapter 4, we defined the rate of profit as the aggregate surplus value divided by the total capital advanced. However, if values have been transformed into production prices, then it is more appropriate to refer to the ratio of aggregate profit to aggregate capital advanced. The expression I/A is the **rate of interest**. It is simply the amount of interest paid (received) divided by the amount of capital borrowed (lent). Finally, the expression R_e/(C+V-A) is the **rate of profit of enterprise**. It is the profit of enterprise divided by the amount of non-borrowed money capital. Finally, k refers to the fraction of the total capital that is borrowed. If we substitute the symbols p for the rate of profit, i for the rate of interest, and r_e for the rate of profit of enterprise, then we have the following result:

$$p = ik + r_e(1-k)$$

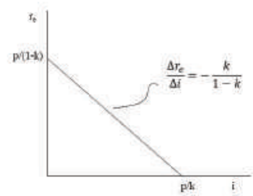
This final expression states that the rate of profit is equal to a weighted average of the rate of interest and the rate of profit of enterprise where k and 1-k serve as the weights.

If we assume that the rate of profit is given, we can then represent all the combinations of the interest rate and the rate of profit of enterprise that are possible in a two-dimensional space. To do so, we solve the above expression for r_e as follows:

$$r_e = \frac{p}{1-k} - \frac{k}{1-k} i$$

This equation turns out to be linear with a vertical intercept of p/(1-k) and a constant slope of –k/(1-k). Given the rate of profit, we can graph the equation as in Figure 15.17.

Figure 15.17: The Relationship between the Rate of Interest (i) and the Rate of Profit of Enterprise (r_e)



Source: Llanos, Theodore. "Marx and the Rate of Interest." *Review of Radical Political Economics*. Volume 19 (3): 34-55, copyright © 1987 by Sage Publications. An adapted version of Figure 1: Reprinted by Permission of Sage Publications, Inc.

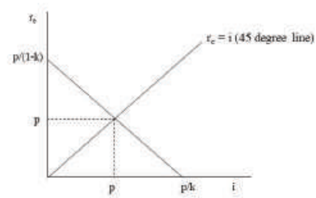
It should be clear that the interest rate and the rate of profit of enterprise are inversely related. That is, as the interest rate rises, the rate of profit of enterprise must fall, and vice versa. This negative relationship is reflected in the downward slope of the line.

More can be stated about the relationship between the three rates that we are considering. For example, consider the point at which the rate of interest and the rate of profit of enterprise are equal (i.e., r_e = i). If we draw a line through all the combinations of r_e and i where these rates are equal, then we obtain a positively sloped line with the equation r_e = i. The reader should note that this line has a slope of 1 and will rise at a 45-degree angle relative to the horizontal axis. The point at which this 45-degree line intersects the line that shows the relationship between i and r_e represents the unique point where r_e and i are equal given this profit rate.

If we set i = r_e and plug it into the equation of the line relating the two variables, then we obtain the following result:
$p = ik + i(1 - k) \Rightarrow p = i$

If p = i, then it necessarily follows that p = r_e as well since i = r_e. All these results are captured in Figure 15.18.

Figure 15.18: The Point at which the Rates of Interest, Profit, and Profit of Enterprise are All Equal



Source: Llanos, Theodore. "Marx and the Rate of Interest." *Review of Radical Political Economics.* Volume 19 (3): 34-55, copyright © 1987 by Sage Publications. An adapted version of Figure 1: Reprinted by Permission of Sage Publications, Inc.

A few interesting results necessarily follow. If the economy is on the downward sloping line at a point above the intersection with the 45-degree line, then the following results must hold:
$r_e > p,\ r_e > i \ and \ p > i$

On the other hand, if the economy is on the downward sloping line at a point below the intersection with the 45-degree line, then the following results must hold:
$r_e < p,\ r_e < i \ and \ p < i$

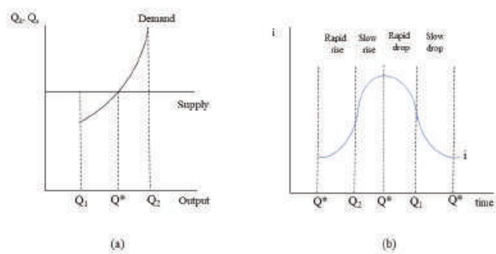
This entire discussion has assumed a constant aggregate rate of profit (p). Over the course of the business cycle, however, the rate of profit fluctuates.[15] Because the vertical and horizontal intercepts of the downward sloping line depend on p, while the slope does not, it follows that parallel shifts to the right (during an expansion) and to the left (during a contraction) will occur over the course of the business cycle.[16]

If the rate of profit fluctuates in a way that can be explained, and the rate of interest fluctuates in a way that can be explained over the course of the business cycle, then fluctuations in the rate of profit of enterprise should be explainable as well using the equation of our downward sloping line.[17] That is, with the movements in two out of three variables explained, the third variable must be explained as well given the logical relationship between the three rates. In Chapter 14, we discussed the fluctuations in the rate of profit according to Marxian theory. Hence, our goal is to explain movements in the rate of interest, which will then complete the explanation of all three rates.

According to Marx, the rate of interest in a capitalist economy is competitively determined in the market for loan capital. This claim sounds rather similar to the neoclassical claim that the loanable funds market plays a central role in the determination of the rate of interest. A key difference exists, however, between Marx's theory and the neoclassical theory. Unlike in neoclassical theory, changes in production are the key to interest rate fluctuations over the course of the business cycle. As output rises and falls, the demand for loan capital changes in response to the needs of industrial capitalists.[18] The degree of tightness in the money capital market then determines whether the interest rate rises and falls as well as the speed of the adjustments.[19]

To clarify the argument, let's assume that the demand for loanable money capital is positively related to the level of production (i.e., output). Let's also assume that the supply of loanable money capital is given or fixed. Both assumptions are reflected in Figure 15.19 (a).[20]

Figure 15.19: The Loan Capital Market and the Movement of the Rate of Interest



Source: Lianos, Theodore. "Marx and the Rate of Interest." *Review of Radical Political Economics*. Volume 19 (3): 34-55, copyright © 1987 by Sage Publications. Adapted versions of Figures 4 and 5: Reprinted by Permission of Sage Publications, Inc.

As output rises during an expansion, beginning at Q1, the quantity supplied exceeds the quantity demanded of loan capital. As a result, the interest rate declines as shown in Figure 15.9 (b). Because the gap becomes smaller between the two curves, the reduction in the interest rate is relatively slow. Once output rises beyond Q*, the demand for loan capital exceeds the supply, and the interest rate rises. Because the gap grows rapidly, the interest rate rises very quickly. After the level of output peaks at Q2, it begins to fall, indicating that a recession has begun. Since the quantity demanded still exceeds the quantity supplied, the rate of interest continues to rise but much more slowly because the gap is becoming smaller. Once output falls below Q*, the interest rate falls very quickly due to the rapid increase in the gap between the supply and demand for loanable funds. Once the trough of the recession is reached at Q_1, it begins to rise once more with a new expansion, and the cycle begins again.[21]

Now that the movement of the rate of interest has been fully explained, this knowledge can be combined with what we know about the movement of the profit rate over the course of the business cycle and the identity that relates the profit rate, the interest rate, and the rate of profit of enterprise. All this information provides an explanation for the movement of the rate of profit of enterprise over the course of the business cycle. With the interest rate following the pattern in Figure 15.19 (b), we can expect the rate of profit of enterprise to fluctuate opposite the interest rate fluctuations given the negative relationship between the two rates.[22]

In closing, it is worth reflecting on the key differences between the Marxian and neoclassical theories of the rate of interest. First, the Marxian theory suggests that the level of production is the most important factor determining movements in the rate of interest, as opposed to the responses of borrowers and lenders to shortages and surpluses in the market for loanable funds.[23] Second, the Marxian theory is a disequilibrium theory of the rate of interest as opposed to its neoclassical counterpart. Even though a point exists (Q^*) where both quantity supplied and quantity demanded are equal in the market for loan capital, the economy has no inherent tendency to move towards this point because the level of production is being driven by other factors.[24] Finally, the rate of interest in Marxian theory is linked to the rate of profit, and the rate of profit reflects how successful capitalists have been in terms of exploiting labor-power. That is, interest income is a portion of the aggregate profit, and aggregate profit is explained on the basis of the surplus labor performed by the working class. By contrast, in neoclassical theory, the rate of interest is the price paid for loanable funds, and the loanable funds market is the market that harmonizes the interests of savers who have a surplus of funds and borrowers who are short of funds. In other words, this market transfers funds from those who lack an efficient use for funds to those who have an efficient use for funds. In short, it is not only the mechanics of interest rate determination that differ across the two theories, it is also the social meaning of the interest rate and the market in which it is determined.

Following the Economic News[25]

Investors were flocking to U.S. government bonds in July 2015. U.S. government bonds are regarded as safe haven assets because the risk of default is so low. The U.S. government has never defaulted on its bonds and so those assets are regarded as an extremely safe investment among investors. The reason for the shift to U.S. government bonds is that U.S. retail sales unexpectedly declined creating a fear among investors that stock prices might fall. As a result, the demand for bonds increased significantly. Bonds are regarded as a substitute for stocks, which are riskier. As we know, a rise in the demand for bonds causes bond prices to rise. As bond prices rise, interest rates fall. In the case of the 10-year Treasury note, the yield fell from 2.43% to 2.401% over a two-day period in July 2015. Similarly, the yield on the two-year Treasury note fell from 0.669% to 0.641%. Because consumer demand weakened, the

Federal Reserve was likely to become even more hesitant about raising short term interest rates, which might have caused the economy to struggle even more. This expectation encouraged investors to buy more bonds, which gave a further boost to bond prices. Investors expected that the Fed would not be engaging in larger bond sales anytime soon. A large volume of bond sales by the Fed would raise supply, lower bond prices, and raise yields. Since the Fed was not expected to act in this way in the very near future, bond prices were not expected to fall, and so investors decided to purchase bonds. The overall result was higher bond prices and lower yields, as predicted by the supply and demand model of the bond market.

Summary of Key Points

1. To calculate the future value (FV) of a sum of money, it is necessary to multiply the present sum by $(1+i)^n$. To calculate the present value (PV) of a sum of money, it is necessary to divide the future sum by $(1+i)^n$.
2. To calculate the present value of a bond, it is necessary to add up the present value of each future payment associated with the bond.
3. Interest rates and bond prices are always inversely related.
4. The demand for loanable funds is equivalent to the supply of bonds, and the supply of loanable funds is equivalent to the demand for bonds.
5. In equilibrium, the price of a bond is equal to its present value.
6. The money market is not the market for short-term securities but the market in which discrepancies between actual money holdings

and desired money holdings influence the rate of interest.

7. The three components of money demand are the transactions demand, the precautionary demand, and the asset demand for money.
8. In the neoclassical model of interest rate determination, competition causes the loanable funds market, the bond market, and the money market to all clear simultaneously, leading to a general equilibrium.
9. Other factors the same, an equilibrium stock price will be lower than an equilibrium bond price due to the greater risk and uncertainty of dividend payments associated with the stock.
10. The real interest rate is calculated as the nominal interest rate minus the rate of inflation.
11. In Austrian theory, a positive rate of interest emerges because consumers have different time preferences and because the problem of time asymmetry exists.
12. In Marxian theory, changes in the level of production lead to changes in the degree of tightness in the loan capital market, which then lead to fluctuations in the rate of interest.

List of Key Terms

Rate of interest

Principal

Compound interest

Future value

Present value

Discounted

Bond

Coupon bonds

Term to maturity

Yield to maturity

Loanable funds market

Equilibrium interest rate

Equilibrium quantity of loanable funds

Bond market

Equilibrium bond price

Equilibrium quantity of bonds

Money market

Money market instruments

Money

Liquidity

Actual money holdings

Demand for money

Transactions demand for money

Precautionary demand for money

Capital gain

Speculative demand for money

Asset demand for money

General equilibrium model

Partial equilibrium model

Marginalist revolution

Walras's Law

Stocks

Underwriting a stock issue

Primary market

Promoters' profit

Dividends

Secondary market

Default

Nominal interest rate

Real interest rate

Purchasing power

Time preference

Positive time preference

Negative time preference

Neutral time preference

Time asymmetry

Profit of enterprise

Rate of profit

Rate of profit of enterprise

Problems for Review

1. Suppose you lend $250 for two years at an annual interest rate of 6.5%. What is the future value of your loan at the end of year 2?
2. Suppose you purchase an asset that will pay you $1000 in three years. If the annual interest rate is 3.5%, then what is the present value of your asset?
3. Suppose you buy a bond that will pay you $300 per year for the next four years. If the annual interest rate is 4%, then what is the price of your bond if the market is in equilibrium?
4. Suppose you purchase a stock that is expected to pay a dividend of $300 per year for the next four years. If the discount rate is 5.5%, then what is the stock price in this case? Is this result the one you would expect when comparing this answer to the one you obtained in question 3? Why or why not?
5. Assume that the loanable funds market, the bond market, and the money market are all in equilibrium. What will happen if the central bank reduces the money supply? Explain what occurs in the money market and how these changes lead to additional changes in the bond and loanable funds markets.
6. Suppose that you lend money at an annual

interest rate of 12% and the annual inflation rate is 7%. What is the real interest rate that you earn?

7. What is the actual net borrowing of a consumer if the actual consumption in a period is 6 apples and the endowment in that same period is 8 apples? What does the sign of your answer imply?
8. Suppose the rate of profit is 50% and the proportion of capital borrowed by industrial capitalists is 75%. Write the equation of the line that relates the rate of profit of enterprise to the rate of interest. Next draw a graph of this line in a space with the rate of profit of enterprise on the vertical axis and the rate of interest on the horizontal axis. Label the horizontal and vertical intercepts as well.

Notes

1. See Mankiw (1997), p. 58, for a justification of references to "the interest rate."
2. Mishkin (2006), p. 105, footnote 4, emphasizes this point.
3. McConnell and Brue (2008), p. 260, present the transactions demand for money as a vertical line. This presentation adds the precautionary demand as a second source of money demand that is independent of the interest rate.
4. For example, Samuelson and Nordhaus (2001), p. 520-521, and McConnell and Brue (2008), p. 259-261, focus on the transactions demand and the asset demand for money only. Mishkin (2006), p. 521-522, on the other hand, concentrates on

the transactions, precautionary, and speculative motives for holding money.

5. Mishkin (2006), p. 143, includes a similar formula for the calculation of a stock price. Mishkin, however, includes a final term representing the final sales price of the stock at the end of the holding period. It is omitted here just as the face value of the coupon bond (i.e., the return of the principal) was omitted in our discussion of how to calculate the present value of a coupon bond.
6. A similar argument can be used to explain the risk structure of interest rates. See Mishkin (2006), p. 120-127.
7. The model presented in this section uses as a starting point the interpretation of Mises's theory of interest as presented in Moss (1978), p. 157-166. Nevertheless, certain liberties have been taken for ease of exposition.
8. Moss (1978), p. 161, attributes the distinction between positive, negative, and neural time preferences to economists Irving Fisher and Gary Becker.
9. See Moss (1978), p. 161, where he states that "an individual satisfied with an equal number of apples in each time period is said to display neutral or zero time preference."
10. See Moss (1978), p. 163-164, for an explanation of the asymmetry in the time market.
11. Moss (1978), p. 160, defines net borrowing as the difference between desired consumption and the endowment.
12. Moss (1978), p. 163, states that "total apple consumption for the entire society must equal total apple endowment when both totals are summed over all individuals and all periods."
13. Moss (1978), p. 163, explains that it is possible "for a single individual's total n-period consumption to be greater or less than his total aggregate apple endowment. Whether it will be

greater or less depends, of course, on whether over the n periods he was a net interest payer or receiver."

14. The interpretation that is discussed in this section was developed by Lianos (1987). Aside from a few notational changes, the mathematical presentation in this section is found in Lianos's original article. The original source for this material is: Lianos, Theodore. "Marx and the Rate of Interest." *Review of Radical Political Economics*. Volume 19 (3): 34-55, copyright © 1987 by Sage Publications. Adapted versions of Figures 1, 4, and 5 with equations 1-6: Reprinted by Permission of Sage Publications, Inc.
15. Lianos (1987), p. 36.
16. Lianos (1987), p. 37, shows these parallel shifts in his Figures 1 and 2.
17. Lianos (1987), p. 38, shows how the rate of profit of enterprise and the rate of interest fluctuate relative to one another in his Figure 3.
18. Lianos (1987), p. 45.
19. Lianos (1987), p. 45-46.
20. The graphs in Figure 15.19 have been re-created using Figures 4 and 5 in Lianos (1987), 46-47.
21. Ibid. p. 46-47.
22. Lianos (1987), p. 37-38, describes the complicated relationship between the two rates throughout the business cycle.
23. As Lianos (1987), p. 52, puts it, "the real sector dominates the monetary sector."
24. Lianos (1987), p. 52, states that the "differences between demand and supply create equilibrating tendencies but they do not lead to equilibrium and stability of the interest rate, because the moving force in this model is the process of accumulation and income growth."

25. Zeng, Min. "U.S. Government Bonds Strengthen." *The Wall Street Journal*. July 14, 2015.

Chapter 16

MONEY SUPPLY MEASURES AND THEORIES OF COMMERCIAL BANK BEHAVIOR

Goals and Objectives:

In this chapter, we will do the following:

1. *Identify* what most economists consider to be the traditional functions of money.
2. *Define* several different types of money
3. *Distinguish* between the Neoclassical and Austrian Measures of the Money Supply
4. *Examine* the difference between 100% reserve banking and fractional reserve banking
5. *Explore* the different items on a commercial bank's balance sheet
6. *Investigate* the neoclassical theory of banks and its link to the theory of financial markets
7. *Develop* a Marxist theory of banks and link it to the theory of financial markets

In Chapter 15, we explored three theories of financial markets. Our focus was on the way in which financial markets determine the rate of interest and cause fluctuations in the rate of interest over time. Our purpose in this chapter is to incorporate the role of money into our neoclassical and heterodox analyses. Therefore, this chapter explores the topic of money, including its functions, major types, and most important measures. How to measure money is somewhat tricky, and neoclassical and Austrian economists do not agree on the best method of measuring its total quantity. This chapter also discusses the banking system, which is closely connected to the subject of money and its amount. A financial statement known as a balance sheet is introduced in this chapter so that it is easier to explore the way commercial banks' activities alter their financial positions and affect the quantity of money in the economy. This background makes it possible to compare two competing theories of how commercial banks operate and how they influence the financial markets. The final part of the chapter, therefore, contrasts the neoclassical and Marxist theories of commercial bank behavior and how those theories connect to the neoclassical and Marxist theories of financial markets discussed in Chapter 15.

The Traditional Functions of Money[1]

When neoclassical economists define money, they do not associate it with any specific object but rather with the functions that it fulfills. In other words, if an object fulfills certain functions, then it is regarded as money. If it does not fulfill these functions, then it is not money. The first function of money is **medium of exchange**. To understand this function, it helps to think about a **barter economy**. A barter economy is an economy in which no money exists and commodities are exchanged for other commodities. For example, five textbooks might exchange for one desk. Barter economies allow owners of commodities to trade for the commodities that they want, but they face one major problem that is referred to as the **double coincidence of wants problem**. For example, suppose that I have five textbooks and would like to purchase one desk. To make this trade, I need to find someone who has a desk to sell. The problem that arises for me, however, is that this person must also want my textbooks if the trade is to occur. For our wants to exactly coincide in this manner would require a double coincidence: the desk owner wants what I have and I want what the desk owner has. If money is introduced into this economy, then the problem is solved. Why? Money by its nature is universally regarded as valuable. Even if the desk owner does not want my textbooks, the desk owner will be pleased to accept my money because it can be used to purchase something else that the desk owner wants, such as a painting, from someone else. That is, money serves as a medium of exchange. It facilitates exchange by making possible the exchange of the desk for the painting, even though this exchange does not occur directly. With money regarded as universally valuable, now I only need to find someone with the commodity that I wish to buy (assuming I have sufficient money to pay the price), and the double coincidence of wants problem is solved.

It should now be clear why neoclassical economists regard money as a means of increasing efficiency. The time required to find someone with whom to make a trade in a barter economy is much greater than the time required in a monetary economy. Because money cuts down on the time required to find someone with whom to make a trade, that freed up time can be used for more productive activities.[2] Also, because of the double coincidence of wants problem, people in barter economies do not want to rely so much on market exchange. It is risky to do so because of the difficulties associated with finding people with whom to trade. In a monetary economy, by contrast, people feel much more comfortable relying on market exchange. As a result, producers specialize more because they know that they can easily sell their products and services for money, which can be used to obtain other commodities they really desire.[3] Specialization and division of labor, of course, greatly increase labor productivity. With greater productivity and overall production, it is easy to see that money makes possible a higher level of efficiency.

A second function that neoclassical economists associate with money is its role as **unit of account**. According to this function, money provides a standardized unit by which to measure prices. That is, money provides a gauge for measuring value. With a standardized unit, it becomes possible to make fine distinctions in the values of commodities. Comparisons of value then become easier to make, which aids decision making. The divisibility of the object is thus of crucial importance. Historically, precious metals like gold and silver, have served as money. Metallic substances can be melted down, weighed, and transformed into standardized units, such as bars and coins. It is important to remember, however, that just because an object can serve as a unit of account, it also must be able to fulfill the other functions of money. If it fails to fulfill the other functions, then societies will reject it for the role of money. For example, milk can be easily measured and careful distinctions in value could be determined if all prices were stated in terms of ounces of milk. Milk does not serve well as a form of money, but it is not because it fails as a unit of account. It fails because it does not fulfill the third function well at all.

The third function that neoclassical economists associate with money is its role as a **store of value**. As a store of value, money makes it possible to transfer purchasing power from the present to the future. To succeed in this role, the object must be highly durable. An object that spoils or corrodes over relatively short time periods will not maintain its value in exchange. Precious metals are highly durable. Milk, on the other hand, spoils after a short time and so if wealth is held in this form, it will quickly vanish and the owner will lose all claim to future commodities.

If all three functions are fulfilled by an object, then neoclassical economists regard the object as money. Societies also tend to use such objects as money because people have recognized, even if not explicitly, the important role that each of these functions serves. Marxian economists also recognize the importance of these functions although the language used is somewhat different (e.g., means of circulation, measure of value, object of hoarding).[4] One important difference between the Marxian and neoclassical analyses of these functions deals with the unit of account (or measure of value) function. Neoclassical economists regard the unit of account function as important because it makes it possible to compare prices of commodities using a common unit. Marxian economists take it further, however, with the argument that money serves as the universal expression of homogeneous human labor. That is, money does not just make it possible to compare the prices of commodities using a common unit, which it certainly does. It also serves as the object through which socially necessary abstract labor time is expressed within the capitalist mode of production. When money serves as a means of circulation (or medium of exchange) as represented in a commodity circuit (C-M-C'), it renders the socially necessary abstract labor time embodied in the two commodities (C and C') commensurable. That is, the embodied SNALT is regarded as equal in the two commodities, and it is regarded as equal to the SNALT embodied in (or reflected in) the money commodity.

Different Types of Money: From Mollusk Shells to Bitcoins

Throughout history many different objects have fulfilled the functions of money to a greater or lesser extent. The form of money that was used for the longest period throughout human history is the cowrie, which is a mollusk shell found in the Indian Ocean.[5] For more than 2,000 years, it served as money at various times in China, India, Europe, and Africa. Many other commodities have circulated as money, including stone wheels, gold, silver, copper, and cigarettes. Each commodity is considered **commodity money** because it circulates as money and yet has some intrinsic value or alternative use. For neoclassical economists, the alternative use of the commodity justifies the label. For Marxian economists, commodity money is also the product of socially necessary abstract labor time, which renders its value in circulation comprehensible.

A more recent form of money is **convertible paper money**. When people refer to the gold standard, for example, they have in mind convertible paper money. During the nineteenth century, for example, commercial banks would issue paper notes that represented claims to their gold reserves. The holder of the paper notes could use the notes to pay bills and settle debts, and the paper was "as good as gold." The U.S. government also printed paper notes that represented claims to gold. Eventually, the U.S. Federal Reserve would issue paper notes that were identified as Federal Reserve Notes, representing the debt of the issuer. That is, the holder possessed a claim to the assets of the central bank. As late as the mid-twentieth century, some U.S. notes were still redeemable for precious metals and so were backed by commodity money. After World War II, gold reserves continued to back U.S. dollars used in international transactions. The international gold standard was abandoned, however, in the 1970s.

If convertible paper money is no longer available, then what kind of money do we have? We now have **inconvertible paper money**. That is, we have paper money that is not convertible into anything else. It cannot be exchanged at any bank for gold, silver, copper, or any other commodity money. Why do people hold money if it is not redeemable into any commodity of value? People hold money because the government has declared it to have value. Because it is declared to be money by government decree, it is labeled **fiat money**. To some extent, the value of fiat money represents people's faith in the government. The more important reason that people hold money, however, is that it is generally accepted in exchange for other commodities. People understand inconvertible paper money to represent a claim to the huge variety of commodities available in the marketplace as well as its acceptance for the payment of debts.

Another type of money that has developed with the rise of information technology is **digital currencies**. The most widely known digital currency is the Bitcoin. Interestingly, Bitcoins are not issued by any central bank. They were introduced in 2009 and are created through a process called **mining**. That is, computer programmers create new Bitcoins by solving mathematical problems. Because the process takes time, the total supply of Bitcoins increases but only gradually. It has no physical existence or paper form. It is purely digital. It has a market value because it can be exchanged for other currencies like the U.S. dollar. Its value has plummeted at times due to low demand and a rising supply and has skyrocketed at other times due to soaring demand and a relatively slow increase in the supply. The existence of privately created monies poses serious problems for governments and central banks. Because the source of Bitcoins in transactions is more difficult to trace in comparison with paper money, they have been used for money laundering and other illegal activities. Furthermore, although Bitcoins are still a relatively new form of money and not nearly as widely used as national currencies, the expansion of its use could one day interfere with the ability of central banks to manage their nations' money supplies. Only time will tell whether digital currencies become so widely used that they become serious competitors for national currencies.

The Neoclassical Approach to Money Supply Measurement

Before discussing how changes in the supply of money affect the economy, it is essential to discuss how it is measured. That is, it is usually helpful to address the issue of measurement before one develops a theory, remembering however that how one chooses to define and measure an economic variable may influence the theory that one develops. We saw in Chapter 1 how this problem arose when we considered how the Bureau of Labor Statistics measures the unemployment rate and the way in which that measure may influence one's perception of a worker without a paid job. Nevertheless, one must begin somewhere, and to begin with the measurement of an economic variable seems like the easier choice.

Neoclassical economists use several primary measures of the **money supply** (also referred to as the **money stock**), and the distinction between the measures deals with the **liquidity** of the assets included in each measure. Liquidity refers to the ease with which an asset can be converted into **currency** (i.e., coins and paper money).

The first neoclassical measure of the money supply includes the most liquid assets and is referred to as the **M1 money supply**. That is, M1 includes publicly held currency and the checkable deposits of commercial banks and thrift institutions. **Thrift institutions** include depository institutions like savings and loan associations and credit unions. **Savings and loan associations** (S&Ls) accept deposits and tend to specialize in home mortgage lending. **Credit unions** also accept deposits and make loans, but their members own and control them. It should be clear why M1 includes the most liquid assets. Publicly held currency is already currency and requires no conversion into currency. Note that currency held in bank vaults is not counted in the M1 money supply. Checkable deposits are also highly liquid assets. Because account holders can write checks against their balances, these deposits are easily accessed for purchases. They are also available for immediate withdrawal in the form of currency. The M1 money supply is the most closely followed measure of the money supply by the Federal Reserve, and it is the measure that is used most widely in economic models given the extreme liquidity of the assets that it includes.

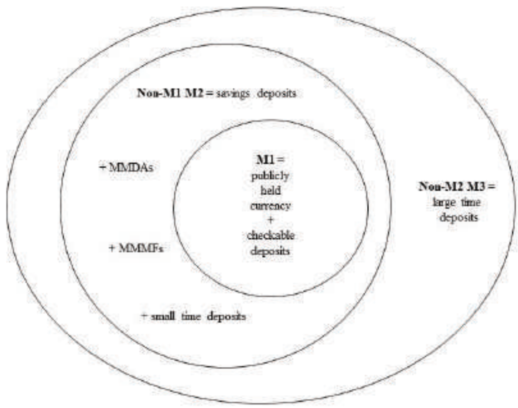
The second neoclassical measure of the money supply includes less liquid assets than M1 and is referred to as the **M2 money supply**. The M2 money supply includes the entire M1 money supply but also includes additional assets that are less easily converted into currency. That is, M2 includes savings deposits, small time deposits (e.g., certificates of deposit valued at less than $100,000), money market deposit accounts (MMDAs), and money market mutual funds (MMMFs).

Savings deposits are held at banks and thrift institutions, but they are less liquid assets than checkable deposits because the check writing feature is absent. In addition, federal law limits the number of withdrawals per month to six. The funds can still be accessed but not quite as easily as the funds in checking accounts. **Certificates of deposit** (CDs) are also less liquid. These financial instruments allow the account owner to deposit a sum of money for a given period to earn a higher interest rate than what can typically be earned on a savings deposit. To earn this higher interest rate, however, the CD owner must commit the funds for a specific period. If the owner withdraws the funds before the maturity date of the CD, then a penalty must be paid. The threat of a penalty and the commitment of the funds for a specific period make this asset less liquid than a checkable deposit or even a savings deposit. **Money market mutual funds** and **money market deposit accounts** appeared in the 1970s and 1980s. Mutual fund companies offer MMMFs to investors who pool their assets together in these funds. The funds are then used to purchase money market instruments, as described in Chapter 15. MMMFs invest in highly liquid short-term assets and so are liquid as well. Investors value them for their low risk, high liquidity, and return that exceeds what can typically be earned on a savings deposit. MMDAs are like MMMFs in that the deposited funds are invested in money market instruments. Banks offer MMDAs, however, and the check writing option is limited, making them less liquid than checkable deposits.

The third neoclassical measure of the money supply is M3. The **M3 money supply** includes M2 but also large time deposits (i.e., CDs with values at least as great as $100,000). These CDs are so large that only institutional investors, like hedge funds and investment firms, can afford them. Because these assets are less liquid due to their longer terms to maturity and considerable penalties for early withdrawal, they are less liquid assets than the components of M2. Overall, the three neoclassical measures of the money supply tend to move together over time, although not always. The divergences between the three measures of the money supply can create a problem for the Federal Reserve as it aims to regulate the money supply.

The diagram in Figure 16.1 helps us to understand the relationship between the three neoclassical measures of the money supply.

Figure 16.1: The Relationship Between the Neoclassical Money Supply Measures



As we move from the M1 money supply to the outer circles, we begin to include less liquid assets and so arrive at the M2 and M3 measures of the money supply. Table 16.1 provides Federal Reserve data on the M1 and M2 money stock measures as of December 2017 so that we can obtain a sense of the magnitudes of the different components of the money supply. We exclude the M3 money supply because the Federal Reserve ceased publication of it on March 23, 2006.[6]

Table 16.1: Estimates of M1 and M2 in December 2017

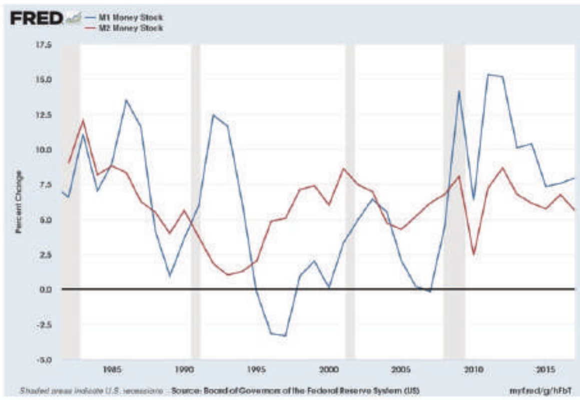
Estimates of the M1 and M2 Money Stock Components, December 2017, seasonally adjusted in billions of dollars	
The M1 Money Supply (M1)	3614.3
Publicly Held Currency (C)	1523.6
Checkable Deposits (D)	2090.7
The M2 Money Supply (M2)	13845
Savings Deposits (S)	9124.5
Certificates of Deposit (CDs)	401.5
Money Market Mutual Funds (MMMFs)	704.8

Source: Board of Governors of the Federal Reserve System. Web. Accessed on January 20, 2018. https://www.federalreserve.gov/releases/h6/current/default.htm
Note: Savings deposits include MMDAs. Traveler's checks have been included in the checkable deposits category. Some rounding prevents the components of M2 from adding to its total in the table.

As Table 16.1 shows, the M1 money supply is relatively small at $3,614.3 billion (or approximately $3.6 trillion) in comparison with the M2 money supply of $13,845 billion, mainly due to the magnitude of the savings deposits. We can also see that currency and checkable deposits make significant contributions to the M1 money supply, but checkable deposits constitute a somewhat larger share of the total.

Figure 16.2 shows the percentage changes in the M1 and M2 money supplies from 1982 to 2017. The two monetary aggregates have frequently moved together, but they sometimes diverge.

Figure 16.2: Growth Rates of the M1 and M2 Money Stocks, 1982-2017



Source: Board of Governors of the Federal Reserve System (US), M1 Money Stock [M1SL] and M2 Money Stock [M2], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/M1SL and https://fred.stlouisfed.org/series/M2, January 20, 2018.

The most obvious example of a divergence in the growth rates of the M1 and M2 money supplies occurred in the mid-1990s. The M1 money stock contracted during this period as reflected in the negative growth rate of M1. At the same time, the growth rate of M2 was positive and rising, indicating that the M2 monetary aggregate was expanding and that the rate of expansion was increasing. The reason for this divergence may relate to the information technology revolution, which made it possible to more easily transfer funds between checkable deposits and non-M1 components of M2 such as savings deposits and money market funds. A similar reduction in the growth rate of M1 while the growth rate of M2 rose occurred in the mid-2000s before the Great Recession. One possibility is that people hold less M1 money during economic booms as they think more about profitable investment strategies. During economic contractions (the shaded areas), on the other hand, M1 money balances appear to increase as people seek security in highly liquid assets.

The Austrian Approach to Money Supply Measurement

Austrian economists agree with some aspects of the neoclassical approach to money supply measurement, but they differ sharply on several key points. Austrian economist Murray Rothbard has written extensively on this subject, and his perspective forms the basis of the Austrian approach described in this section.

First, Austrian economists strongly reject the notion that the liquidity of an asset should determine whether it is considered a part of the money supply. According to Rothbard, "the money supply should be defined as all entities which are redeemable on demand in standard cash at a fixed rate."[7] Obviously this definition includes currency held by the public. It also includes demand deposits. Even though banks generally do not have sufficient cash reserves to redeem all outstanding demand deposits, Austrian economists emphasize the central role of subjective estimates of value by market participants. Therefore, if individuals believe that demand deposits are redeemable, they represent an active part of the money supply.[8] Austrian economists regard savings deposits held at commercial banks and savings and loan associations to be part of the money supply as well. Although some restrictions on transactions apply, depositors subjectively treat their savings deposits as redeemable for cash at a fixed rate and hence, they should be treated as part of the supply of money.[9]

Time deposits (CDs), however, are treated rather differently in the Austrian definition of the money supply. Rothbard draws upon the theory of money and credit that Ludwig von Mises developed to make this argument. Whereas a demand deposit represents a claim to cash and can be used in the purchase of present goods, a time deposit represents a credit instrument and can only be used for the purchase of future goods.[10] One cannot argue that time deposits are highly liquid assets and so should be counted as part of the money supply. Such assets are sold at market rates and are not directly redeemable for cash.[11] On the other hand, CDs and federal savings bonds are redeemable at fixed, penalty rates. Therefore, Austrian economists believe that we should include these values at their penalty levels as opposed to their face values (e.g., $9,000 as opposed to $10,000 when early withdrawal occurs).[12]

Three other issues must be discussed before providing the complete Austrian definition of the money supply. First, consistent with the Austrian criteria, the **cash surrender values** of life insurance policies must be included in the money supply because such policies are redeemable in cash. It is necessary, however, to add the total policy reserves minus the policy loans outstanding because the policy loans are not available for immediate withdrawal.[13] Second, if noncommercial banking institutions, such as life insurance companies or savings and loan associations, have deposits that act as reserves supporting their own issued deposits, then to avoid double counting, those reserves must be subtracted from the total demand deposits when calculating the money supply.[14] Finally, it is essential to include U.S. Treasury deposits held at the Federal Reserve in the money supply because such deposits may be used for the purchase of present goods.[15]

We can now state the complete **Austrian definition of the money supply** (M_a), which contrasts with the neoclassical definitions of the money supply (M1, M2, and M3).

M_a[16] = the total supply of cash

-cash held in the banks

+total demand deposits (including Treasury deposits)

+total savings deposits in commercial and savings banks

+total shares in savings and loan associations (which function like savings deposits)

+time deposits and small CDs at current redemption rates

+total policy reserves of life insurance companies

-policy loans outstanding

-demand deposits owned by savings banks, savings and loan associations, and life insurance companies

+savings bonds, at current rates of redemption

The first three items in the formula, relating to currency and demand deposits, resemble the components of M1. Savings deposits and shares closely resemble the non-M1 elements of M2. It is in the treatment of time deposits, cash surrender values, demand deposits of thrift institutions, and federal savings bonds where we see the greatest difference between the Austrian definition of the money supply and the neoclassical definitions of the money supply. The major reason for the differences in the definitions is that neoclassical economists concentrate on the liquidity of the assets in their definitions whereas Austrian economists focus on the potential to redeem the assets for cash at fixed rates even if the redemption rates are penalty rates.

The Origin of Fractional Reserve Banking

Now that we understand the most important measures of the money supply, we will begin thinking about how the money supply is determined within capitalist economies. We thus move from the practical problem of measurement to the theoretical problem of determination. To address the problem of money supply determination, we will consider two contrasting theories of the way in which a private commercial banking system helps determine the M1 money supply. Because the central bank also plays a major role in determining the M1 money supply, the neoclassical and Marxist theories of money supply determination that we develop in this chapter are only partial theories. The role that the central bank plays in determining the money supply is the subject of Chapter 17.

To understand the contribution of private commercial banks to the determination of the M1 money supply in neoclassical theory and in Marxist theory, we need to first distinguish between 100% reserve banking systems and fractional reserve banking systems. To make sense of this distinction, it is helpful to imagine an economy with a money commodity like gold.[17] Suppose that owners of gold bars decide that it is a great burden to use gold bars each time they wish to buy commodities. Some enterprising young person sees this problem as an opportunity to earn an income. Suppose that she offers to safeguard the gold bars for their owners (for a fee) and offers to issue paper certificates in an amount that is equivalent to the gold bars. These paper certificates are payable in gold on demand. The entire money supply now consists of the total supply of paper certificates, plus any gold bars that remain in circulation. Although a money commodity may still circulate, convertible paper money is now the dominant type of money in this economy.

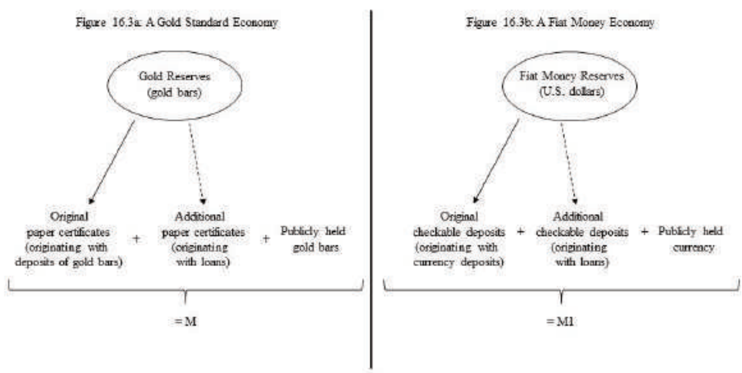
This economy possesses what economists call a **100% reserve banking system**. That is, 100% of the circulating paper certificates are backed up with gold reserves. If every holder of a paper certificate decided at the same time to redeem their certificates for gold, then the banker who issued the certificates would be able to satisfy every certificate owner's demand for gold bars. The likelihood of everyone redeeming their paper certificates at once is very low, but it could happen if the holders of the certificates begin to doubt whether the banker will pay them in gold. If their faith is shaken enough, then a **bank run** might occur where every certificate owner demands payment in gold at the same time. It is worth noting that bank runs of this kind create no problems in a 100% reserve banking system because the demand for gold bars can be completely satisfied without delay.

If sufficient faith in the 100% reserve banking system exists, then most holders of paper certificates are very unlikely to redeem their paper notes for gold. After all, the convertible paper money was created so that owners of gold bars would not need to concern themselves with the protection of the gold bars or the burden of hauling them around to engage in market exchanges. It is the low rate of redemption of the paper certificates that causes the private banker to consider another potential source of income. She decides to create and issue new paper certificates to the point where the face value of the certificates exceeds the quantity of gold reserves that she holds. By doing so, she has expanded the convertible paper money supply and has thus created new money. Rather than using these new certificates for purchases, which she could do, the private banker decides to grant loans, in the form of paper certificates, to borrowers who agree to repay her at a future date with interest. The granting of loans has thus expanded the paper money supply. The total money supply now consists of the original certificates that were issued in exchange for deposits of gold bars, the newly issued certificates representing loans to borrowers, and any gold bars that remain in circulation. The total money supply (M) can be calculated as follows:

M = original certificates + additional certificates (representing loans) + publicly held gold bars

Because the face value of the convertible paper money supply exceeds the value of the gold reserves held in the bank, the banking system is regarded as a **fractional reserve banking system**. That is, the reserves held in the bank only partially support the convertible paper money that is circulating in the economy. In other words, the value of the reserves only equals a fraction of the value of the circulating paper money. Now a bank run has serious consequences for the banker. If every holder of a paper certificate decides to redeem the certificates for gold bars, then the banker will not be able to satisfy those demands. If the rule of law is not firmly established, then the banker might suffer violence at the hands of an angry mob. Figure 16.3a summarizes the results of this analysis:

Figure 16.3: Fractional Reserve Banking in Gold Standard Economies and Fiat Money Economies



From an efficiency perspective, neoclassical economists argue that it is understandable that institutions would create and issue convertible paper money. Holders of gold and silver do not want to pay for other commodities directly using these commodities because they are heavy and difficult to transport. Banking institutions arose to specialize in the safeguarding of these assets. Owners of gold and silver are willing to pay for this service and bankers aim to profit from its provision. Nevertheless, it is the granting of loans and the creation of new money that has the potential to make banking so profitable and so risky because of the constant danger of bank panics and bank runs.

Because modern economies have abandoned commodity money, you might think that this analysis is obsolete with no practical application to modern banking systems. On the contrary, this analysis is entirely applicable to modern banking systems. The only important difference is that the material that serves as bank reserves and the asset that serves as a claim to those reserves have changed.

The dominant type of money in the U.S. economy is government-issued fiat money, as described earlier in this chapter. Although paper money is easier to use in transactions than gold bars, it is still a burden to use for all transactions, particularly large transactions. Therefore, commercial banks accept currency deposits, which perfectly parallels the deposits of gold bars in our economy based on the gold standard. It would not make much sense to issue paper certificates in exchange for the currency deposits, however, because paper would then simply circulate in place of paper. Instead, banks issue checkable deposits when currency deposits are made. The checkable deposits are analogous to the paper certificates in the gold standard economy. Purchases are made using paper checks that can be written in any amount to access the underlying deposits. It is important to remember that it is the underlying checkable deposits that represent money in modern banking systems, not the paper checks that only allow one to access them. The total money supply (M1) for this economy can be calculated as follows:

M1 = original checkable deposits + additional deposits (representing loans) + publicly held currency

Just as in the case of the gold standard economy, private commercial banks recognize that they can issue checkable deposits in an amount that exceeds the reserves of U.S. dollars that they hold. It is a desire to grant interest-bearing loans that motivates the issuance of additional checkable deposits. Because checkable deposits are part of the M1 money supply, the issuance of additional checkable deposits represents the creation of new money. A 100% reserve banking system is possible in an economy based on fiat money if the banks only create checkable deposits in response to the receipt of currency deposits. As soon as the face value of the deposits exceeds the reserves of U.S. dollars held in bank vaults, the system becomes a fractional reserve banking system and is vulnerable to the threat of bank runs. Figure 16.3b summarizes the results of this analysis.

The Balance Sheet of a Commercial Bank

To investigate more carefully how commercial banks influence the total quantity of money available in the economy, it is helpful to use a financial statement known as a balance sheet. A **balance sheet** is a financial statement that provides a snapshot of an entity's financial position at a point in time. Balance sheets can be constructed for individuals and for firms. The balance sheet has two columns. On the left side of the balance sheet, all the assets of the individual or firm are listed with their associated values. The **assets** are simply items of property that the entity owns. The right side of the balance sheet lists all the liabilities and the net worth of the entity. Liabilities and net worth represent claims against the assets. **Liabilities** are external claims to the assets of the entity. That is, they represent the claims of non-owners to the entity's assets and so they are debts for the individual or firm. **Net worth** (also referred to as equity) represents an internal claim to the assets of the household or firm. That is, after accounting for all the debts, the net worth of the entity represents its claim to the remaining assets. When people sometimes ask how much someone is worth, they have in mind their net worth. That is, they are asking how much property a person owns after subtracting her debts. We can obtain a better understanding of the balance sheet concept if we consider a simple example of how this financial statement might look for one individual as shown in Figure 16.4.

Figure 16.4: The Balance Sheet of an Individual

Assets	Liabilities + Net Worth
Cash $200	Auto loan $5000
Checkable deposit $500	Credit card loan $2000
Savings deposit $1000	Net worth $3,700
Automobile $9,000	
TOTAL = $10,700	TOTAL = $10,700

The individual whose balance sheet is represented in Figure 16.4 possesses several different types of assets, including cash, a checkable deposit, a savings deposit, and an automobile. She also has incurred debts to obtain these assets, including an auto loan and a credit card loan. The auto and credit card loans represent liabilities because the lending agencies have claims to the individual's assets.

Because the individual's assets exceed her liabilities, her net worth is positive, representing her claim to her own assets. Because every asset will be claimed by someone (i.e., owners or non-owners), the total assets must equal the sum of the liabilities and net worth. The **balance sheet equation** may thus be written as:

Assets = Liabilities + Net Worth (equity)

It should be noted that liabilities may exceed assets. When this situation arises, net worth must be negative for the balance sheet equation to hold. A negative net worth means that the individual or firm is **insolvent** and may need to declare bankruptcy to rectify the situation.

In this chapter, we are primarily interested in how the activities of private commercial banks influence the money supply and the financial markets. Therefore, we need to look at the balance sheet of a commercial bank. The items that typically appear on a bank's balance sheet are different from the items that generally appear on a firm's balance sheet or an individual's balance sheet. Figure 16.5 shows a consolidated balance sheet for all commercial banks in the United States as of January 10, 2018.[18]

Figure 16.5: The Consolidated Balance Sheet of all U.S. Commercial Banks, January 10, 2018, seasonally adjusted, billions of U.S. dollars

Assets	Liabilities + Bank Capital
Securities $3,447.6	Deposits $11,960.2
Loans $9,113.5	Borrowings $2,093.6
Loan losses ($110)	Other $817.4
Cash reserves $2,405.4	Bank capital $1,870.3
Other $1,885	
TOTAL = $16,741.5	TOTAL = $16,741.5

Source: Board of Governors of the Federal Reserve System (U.S.), H.8 Assets and Liabilities of Commercial Banks in the United States. Web. Release Date: 1/19/2018. Retrieved on 1/22/2018. http://www.federalreserve.gov/releases/h8/current/

Note: Rounding error exists in the calculation of the total assets. The total assets were reported as $16,741.4 billion, but the figure has been adjusted for illustrative purposes to match the total on the right hand side.

Only the most essential balance sheet items have been included in Figure 16.5. The asset side of the balance sheet includes **securities**, which totaled nearly $3.5 trillion in January 2018. Commercial banks purchase income-earning assets such as U.S. Treasury securities and mortgage-backed securities (MBSs). U.S. Treasury securities include short-term money market instruments such as Treasury bills that have terms to maturity of less than one year and long-term capital market instruments such as Treasury notes and Treasury bonds that have terms to maturity of greater than one year. Mortgage-backed securities are assets that represent bundles of many mortgage loans with varying degrees of risk that generate interest income for the owner over time. MBSs were created rapidly during the leadup to the 2008 financial crisis and banks continue to hold large quantities of these assets.

Loans are another important asset on the consolidated balance sheet of U.S. commercial banks. In fact, loan assets are the largest item on the balance sheets of banks at over $9.1 trillion in January 2018. This fact should not surprise the reader. After all, fractional reserve banking developed precisely because banks discovered that they could earn interest income by granting loans to borrowers. This category includes loans of all kinds, including commercial loans, industrial loans, real estate loans, and consumer loans. It is the most profitable activity in which banks engage, but it is also the riskiest activity because borrowers sometimes **default** (or fail to repay) their loans with interest. When such defaults occur, banks experience **loan losses**, which accounts for the $110 billion asset reduction in Figure 16.5. During the Great Recession, a wave of defaults on residential mortgages occurred, which caused the values of MBSs to collapse and wiped out a massive amount of wealth that only existed on paper. If such loan losses are large enough, then it might cause assets to fall below liabilities, causing insolvency.

The final item of importance on the asset side of commercial banks' balance sheets is **cash reserves**. Cash reserves include currency that is held in bank vaults, which is not part of M1. It also includes funds that are held in accounts with the Federal Reserve. Cash reserves are important because they make it possible for banks to satisfy depositors' demands for cash withdrawals or withdraws that occur by means of writing checks. Within a fractional reserve banking system, these cash reserves are generally a fraction of the total deposit liabilities. In this case, cash reserves are approximately 20.11% of total deposit liabilities (= 2,405.4/11,960.2) so that for each dollar of deposits, the bank maintains about $0.20 in cash reserves.

On the liabilities and net worth side of the balance sheet, the main categories include deposits and borrowings. **Deposits** are the largest category of liabilities at nearly $12 trillion in January 2018. Deposits are created when depositors transfer funds to banks, either by means of cash deposits or check writing. Because depositors do not own the bank, their claims to the assets of banks are external claims or liabilities. This category of liabilities includes checkable deposits but also savings deposits and time deposits.

The other major category of liabilities for banks is borrowings at just over $2 trillion. **Borrowings** include funds borrowed from other commercial banks, the Federal Reserve, and corporations. When commercial banks borrow from other commercial banks, they do so in the federal funds market. The **federal funds market** is the market for overnight loans made between banks when they lend their reserves to one another. The interest rate that emerges in this market due to the competitive interaction between borrowing banks and lending banks is the **federal funds rate**. This interest rate fluctuates with changes in the supply and demand for reserves. Another type of borrowing is borrowing from the Federal Reserve. Loans that the Federal Reserve grants to commercial banks are called **discount loans**. The Fed charges an interest rate on such loans that is referred to as the **discount rate**. Finally, banks also borrow from corporations. Banks sell large certificates of deposit to corporations in exchange for the use of funds for a fixed period. In exchange for this privilege, banks pay high rates of interest to the lenders.

The final item of importance on the banks' balance sheet is what we have previously called net worth or equity.

The owners of the banks possess a claim to the assets of banks. In this case, the owners are the stockholders, and so net worth is sometimes referred to as stockholders' equity. In the special case of banking institutions, net worth is also referred to as **bank capital**. Because the phrase is so widely used, when we refer to the net worth of banks, we will refer to bank capital. In January 2018, total bank capital for U.S. commercial banks approached \$1.9 trillion. Adding together the items on each side of the balance sheet yields approximately \$16.74 trillion of assets and \$16.74 trillion of liabilities plus bank capital. As expected, the two sides of the balance sheet add up to the same total. Although we have considered the consolidated balance sheet of all U.S. commercial banks, each commercial bank has its own balance sheet. The balance sheet equation holds for each bank's balance sheet, and each item on the consolidated balance sheet is simply the total for that item across all individual banks' balance sheets.

A Neoclassical Theory of Commercial Banks and its Theory of Financial Markets

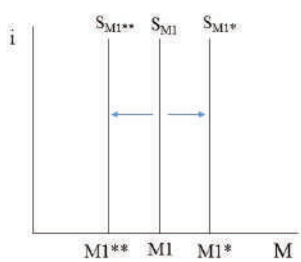
This section explores the way in which neoclassical economists interpret commercial bank behavior and how banks influence the total stock of money. The neoclassical analysis that we develop is also linked to the neoclassical theory of financial markets presented in Chapter 15. That is, it will be shown how neoclassical theory provides an explanation for movements in bond prices and interest rates.

The reader should keep in mind that commercial banks are profit-seeking institutions just like production firms. They do not aim to alter the supply of money but do so when seeking maximum profits. To maximize profits, banks must undertake a delicate balancing act that involves carefully adjusting the various items on their balance sheets. Indeed, the profits of a bank depend on the selected mix of the different conflicting balance sheet items.

The management of banks involves different and often competing objectives.[19] For example, banks must purchase assets that will generate maximum income for the bank. Loan assets are the greatest generator of bank income, but securities are also an important source of income for banks. At the same time, banks must also acquire funds at minimum cost to the bank. Because liabilities include deposits of all kinds, banks must pay interest to acquire and maintain them. Even in the case of checkable deposits that pay no interest at all, commercial banks compete for new depositors and frequently offer cash bonuses to depositors who open new checking accounts. Banks also need to acquire and maintain sufficiently liquid assets so that the banks can satisfy depositors' demands for withdrawals. Cash reserves are the most liquid bank asset. Short-term securities, such as Treasury bills, are also relatively liquid assets and can be converted into currency at relatively low cost. Finally, banks are required to maintain sufficient levels of bank capital so that if loan losses occur, the banks will not become insolvent. Since liabilities are fixed, when loan losses occur, net worth or bank capital must shrink to maintain overall balance on the balance sheet. If bank capital falls to the point where it becomes negative, then the bank is technically insolvent.

The way that private commercial banks strategize to achieve maximum profits is a subject of great interest, but we set it aside in this chapter to concentrate on the influence that banks and private individuals have on the major measures of the money supply. In other words, we are interested in the reasons for shifts in the money supply curve that occur in the money market as described in Chapter 15. Figure 16.6 shows how the M1 money supply may increase to M1* or decrease to M1** in the money market.

Figure 16.6: Changes in the M1 Money Supply in the Money Market



To explore the role of banks in changing the money supply, we only show the changes that occur to the balance sheet items and exclude all items that remain unaffected when a transaction occurs. This approach greatly simplifies the presentation and allows us to focus on the consequences of interactions between commercial banks and other entities. The impact on the monetary aggregates should also be clearer.

The first type of transaction that we will consider involves cash deposits and cash withdrawals. Suppose that a cash deposit of $500 is made in a bank. Figure 16.7 (a) shows the changes to the balance sheet that result from the deposit.

Figure 16.7: Cash Deposits and Cash Withdrawals Involving Checkable Deposits

Figure 16.7 (a): Cash Deposits in a Checking Account

Assets	Liabilities + Bank Capital
Vault Cash +$500	Checkable Deposits +$500

Figure 16.7 (b): Cash Withdrawals from a Checking Account

Assets	Liabilities + Bank Capital
Vault Cash -$500	Checkable Deposits -$500

When a $500 cash deposit occurs, the bank acquires $500 in cash reserves and $500 in deposit liabilities. The transaction has no impact on the M1 money supply. Although checkable deposits increase, publicly held currency falls by an equal amount. The overall result is an unchanged M1 money supply. M2 is similarly unaffected because M2 includes M1. With M1 remaining the same, M2 remains the same as well.

$$\overline{M1} = C\downarrow + D\uparrow$$

$$\overline{M2} = C\downarrow + D\uparrow + nonM1\ M2$$

Now consider a cash withdrawal of $500. This transaction also has no impact on the M1 or M2 money supplies. When the withdrawal occurs, deposit liabilities fall by the same amount as publicly held currency rises. M1 and M2 are thus unchanged in this case. Figure 16.7 (b) shows the changes to the balance sheet that occur when a $500 cash withdrawal occurs.

$\overline{M1} = C\uparrow + D\downarrow$

$\overline{M2} = C\uparrow + D\downarrow + nonM1\ M2$

Similar changes to the balance sheet occur when these transactions are carried out by check rather than with currency. For example, when an account holder deposits a $500 check in Bank A, the check is sent to the Federal Reserve Bank in that district. The Fed then credits the account of Bank A, thus increasing its reserves. Deposit liabilities also increase for Bank A. The only difference is that a cash deposit has an immediate impact on vault cash, whereas the deposit of a check has a quick, although not immediate, impact on the bank's reserves held at the Fed. The impact is not immediate because the check must be processed and must clear before the bank's reserves increase in the case of a check being deposited.

If the check was written against an account at Bank B, then Bank B experiences a loss of reserves when the Fed adjusts the reserves. Bank B also experiences a reduction in its deposit liabilities. Because deposits at Bank A increase and deposits at Bank B decrease, the total M1 and M2 money supplies remain the same.

$\overline{M1} = C + D\uparrow\downarrow$

$\overline{M2} = C + D\uparrow\downarrow + nonM1\ M2$

Figure 16.8 shows the impact on the Federal Reserve's balance sheet.

Figure 16.8: The Role of the Fed in Processing a Check

The Federal Reserve's balance sheet adjustments: A check is deposited in Bank A, which is drawn against an account at Bank B. Both banks have accounts with the Federal Reserve.

Assets	Liabilities + Bank Capital
	Reserves of Bank A +$500
	Reserves of Bank B -$500

Figure 16.8 shows how the Fed's balance sheet adjusts due to the processing of the check. The changes illustrate a couple important points. First, the reserves that banks hold in their accounts with the Fed are treated as liabilities of the Fed even though they are assets for the banks. The Fed owes the banks these funds, and they represent external claims to the Fed's assets. Second, balance sheet changes need not affect both sides of a balance sheet. That is, an increase in one category of liabilities may be offset by a reduction in another category of liabilities. The Fed's total liabilities do not change, but these changes make it possible to increase the reserves of Bank A and to reduce the reserves of Bank B. Those changes cause the deposit liabilities of Bank A to increase and the deposit liabilities of bank B to decrease. Because the total checkable deposit liabilities in the banking system remain the same, the M1 and M2 money supplies remain unchanged.

Although cash deposits and withdrawals involving checking accounts have no impact on the M1 and M2 money supplies, the result is rather different when we begin looking at cash deposits and withdrawals involving savings accounts. Suppose that a depositor makes a $500 cash deposit in a savings account. The result is shown in Figure 16.9 (a).

Figure 16.9: Cash Deposits and Cash Withdrawals Involving Savings Deposits

Figure 16.9 (a): Cash Deposits in a Savings Account

Assets	Liabilities + Bank Capital
Vault Cash +$500	Savings Deposits +$500

Figure 16.9 (b): Cash Withdrawals from a Savings Account

Assets	Liabilities + Bank Capital
Vault Cash -$500	Savings Deposits -$500

The cash deposit causes savings deposit liabilities to increase and publicly held currency to fall. Because publicly held currency falls with no change in *checkable* deposits, the result is a decrease in the M1 money supply. The M2 money supply remains unaffected, however, because savings deposits are included in M2, as is publicly held currency. In other words, the rise in savings deposits exactly offsets the decline in publicly held currency in M2.

$$M1 \downarrow = C \downarrow + D$$

$$\overline{M2} = C \downarrow + D + nonM1\ M2 \uparrow$$

It is the reduction in M1 that is of interest, however, because it suggests that the ordinary operation of accepting cash deposits can affect the M1 money supply when the deposits are made in savings accounts. Therefore, the public can influence the M1 money supply with its decision to deposit cash in savings accounts. Similarly, cash withdrawals from savings accounts increase the M1 money supply and leave M2 unchanged. In that case, publicly held currency increases as savings deposits fall.

$$M1 \uparrow = C \uparrow + D$$

$$\overline{M2} = C \uparrow + D + nonM1\ M2 \downarrow$$

Figure 16.9 (b) shows the impact of a cash withdrawal of \$500 from a savings account on the bank's balance sheet.

A deposit of a \$500 check drawn against a checkable deposit at one bank into a savings account at another bank would also decrease the M1 money supply. In that case, checkable deposits at one bank would fall and savings deposits at another bank would rise. M1 would fall and M2 would remain the same.

$$M1 \downarrow = C + D \downarrow$$

$$\overline{M2} = C + D \downarrow + nonM1\ M2 \uparrow$$

The Fed would adjust the reserves of the two banks, as shown in Figure 16.8, but the impact on the M1 and M2 money supplies is the same as what we observe in the case of a cash deposit in a savings account.

One other transaction worth considering is a transfer between accounts. If an individual owns a checkable deposit and a savings deposit and decides to transfer $500 worth of funds from her checking account to her savings account, then the M1 money supply falls by $500 and the M2 money supply remains the same.

$$M1 \downarrow = C + D \downarrow$$

$$\overline{M2} = C + D \downarrow + nonM1\ M2 \uparrow$$

Again, checkable deposits are included in M1 but not in M2. The resulting change to the bank's balance sheet due to this transfer is shown in Figure 16.10 (a). Similarly, a transfer of $500 from a savings account to a checking account increases the M1 money supply by $500 and leaves M2 unchanged.

$$M1 \uparrow = C + D \uparrow$$

$$\overline{M2} = C + D \uparrow + nonM1\ M2 \downarrow$$

The impact of a transfer from a savings deposit to a checkable deposit on a bank's balance sheet is shown in Figure 16.10 (b).

Figure 16.10: Transfers between Savings Deposits and Checkable Deposits

Figure 16.10 (a): A Transfer from a Checking Account to a Savings Account

Assets	Liabilities + Bank Capital
	Checkable Deposits -$500 Savings Deposits +$500

Figure 16.10 (b): A Transfer from a Savings Account to a Checking Account

Assets	Liabilities + Bank Capital
	Checkable Deposits +$500 Savings Deposits -$500

This section has analyzed the impact on M1 of movements of funds between publicly held currency or checkable deposits, and savings deposits. The same results apply to similar movements of funds between publicly held currency or checkable deposits into and out of other non-M1 aspects of M2, including MMDAs, MMMFs, and time deposits.

Members of the public can certainly have an impact on the M1 money supply when they make deposits into M1 from savings accounts and the other non-M1 aspects of M2. Withdrawals from M1 into savings accounts and the other non-M1 aspects of M2 also influence M1. The more important source of changes in the M1 money supply, however, include activities that banks initiate. For example, when loans are granted or are repaid, they have a direct impact on the M1 and M2 money supplies.

First consider what happens when a commercial bank grants a $3,000 loan to a borrower. The borrower signs a loan contract or a promissory note, agreeing to repay the loan within a specified period with regular interest payments. The bank accepts the promissory note, which is a non-money asset and credits the checking account of the borrower, which is an M1 money asset. The impact of the transaction on the balance sheet of the bank is represented in Figure 16.11.

Figure 16.11: The Expansion of the M1 Money Supply through Lending

Assets	Liabilities + Bank Capital
Loan Asset +$3,000	Checkable Deposits +$3,000

The bank has acquired a loan asset and has created a checkable deposit. It should be noted that the bank has no additional cash reserves and so publicly held currency did not change. The deposit was created using a non-money asset. The bank has created new money! Both the M1 and M2 money supplies increase due to the loan.

$$M1\uparrow = C + D\uparrow$$

$$M2\uparrow = C + D\uparrow + nonM1\ M2$$

The creation of new money with the issue of additional checkable deposits parallels the issue of new paper certificates in the gold standard economy we considered earlier in this chapter. The bank sees an opportunity to make a profit and takes advantage of the fact that most deposits will not be withdrawn during the period of the loan. The borrower's deposit is likely to be quickly withdrawn, however, because the borrower has incurred considerable expense to obtain the funds and plans to use them. Whether it is withdrawn in the form of currency or due to the Fed's check processing activities, the M1 money supply is not expected to be directly affected due to the withdrawal. It is the granting of the initial loan, however, that has created new money.

The bank incurs some risk when it creates new money because the more money it creates, the more likely it is that it will not have sufficient reserves to satisfy depositors who write checks against their balances or who make cash withdrawals. Therefore, the Federal Reserve imposes a **reserve requirement**, which is a legal requirement that banks maintain a minimum fraction of their deposit liabilities as reserves, which is stated as a percentage of the bank's deposit liabilities. Banks may also choose to hold **excess reserves** above the reserve requirement. To understand how reserve requirements work, consider the balance sheet of the bank shown in Figure 16.12.

Figure 16.12: A Bank Subject to a 10% Reserve Requirement

Assets	Liabilities + Bank Capital
Securities $35,000	Deposits $100,000
Loans $90,000	Borrowings $22,000
Cash reserves $12,000	Bank capital $15,000
TOTAL = $137,000	TOTAL = $137,000

The bank is subject to a 10% legal reserve requirement. It has cash reserves of $12,000 and deposit liabilities of $100,000. Clearly, the banking system is a fractional reserve banking system since not all deposits are supported with fiat money reserves. Nevertheless, this bank is meeting its reserve requirement since the ratio of currency to deposits (C/D) is equal to 12%. In fact, the bank is exceeding the minimum requirement and has excess reserves of $2,000, which could be used to legally expand the money supply through lending.

To see the impact on the balance sheet of a bank due to the repayment of a loan, let's begin with Figure 16.3a, which shows what happens when the borrower withdraws the $3,000 loan that was granted in Figure 16.11.

Figure 16.13a: The Withdrawal of the Loan Amount

Assets	Liabilities + Bank Capital
Reserves -$3,000	Checkable Deposits -$3,000

Figure 16.3a shows that the bank's deposit liabilities decline by $3,000, but its reserves also decrease by $3,000. The borrower has either withdrawn the funds in the form of currency or has written a check against the amount. During the period of the loan, the loan asset increases in value due to the anticipation of a final payment of interest when the loan matures. If the interest rate (i) is 5%, then the loan asset appreciates by $150 (=5% times $3,000). Because the bank's deposit liabilities have not changed, the bank's shareholders enjoy an increase in bank capital. This change is represented in Figure 16.3b.

Figure 16.13b: The Growth of the Loan Asset with the Accumulation of Anticipated Interest (i = 5%)

Assets	Liabilities + Bank Capital
Loan Asset +$150	Bank Capital +$150

When the loan matures, the borrower must repay the principal amount plus interest. She, therefore, deposits the full amount of $3,150 in her account at the bank. The bank's deposit liabilities rise by this amount. The bank's reserves also increase by this amount giving balance to the balance sheet. This change is shown in Figure 16.3c.

Figure 16.13c: The Borrower Deposits Funds to Repay the Interest and Principal Amount of the Loan

Assets	Liabilities + Bank Capital
Reserves $3,150	Checkable Deposits $3,150

Finally, the borrower repays the loan with a check drawn against her account at the bank. The bank debits the borrower's account, which reduces the bank's deposit liabilities. It also returns the canceled promissory note to the borrower as shown in Figure 16.13d.

Figure 16.13d: The Contraction of the M1 Money Supply through Loan Repayment

Assets	Liabilities + Bank Capital
Loan Asset -$3,150	Checkable Deposits -$3,150

Notice that the bank has $3,150 in reserves, which is $150 more than the amount with which it began. It also has increased its bank capital by $150, which reveals how interest-bearing loans serve as a means of appropriating profits for the bank and enhancing bank capital. Most importantly, we see that when loans are repaid, the M1 and M2 money supplies contract. Borrowers repay loans with interest. They receive the canceled promissory notes, and their checking accounts are debited. A non-money asset returns to circulation, and a money asset leaves circulation.

It is assumed that the borrower redeposited the initial loan amount and made an additional cash deposit of $150 to make possible the repayment of interest with the loan. When these cash deposits are made, the M1 and M2 money supplies do not change because publicly held currency falls as much as checkable deposits increase. It is only when the loan is repaid that the M1 money supply falls by the full amount of the repayment. Interest payments on bank loans thus contribute to the contraction of the money supply.

In Figure 16.13d, the M1 and M2 money supplies fall because checkable deposits decline with no change in publicly held currency.

$$M1\downarrow = C + D\downarrow$$

$$M2\downarrow = C + D\downarrow + nonM1\ M2$$

It is the canceled promissory note that returns to the borrower, not currency, as would occur in the case of a cash withdrawal. Hence, both monetary aggregates contract. It is worth noting that banks can change both M1 and M2 with their lending activities whereas deposits and withdrawals from savings accounts (and other non-M1 components of M2) only affect M1.

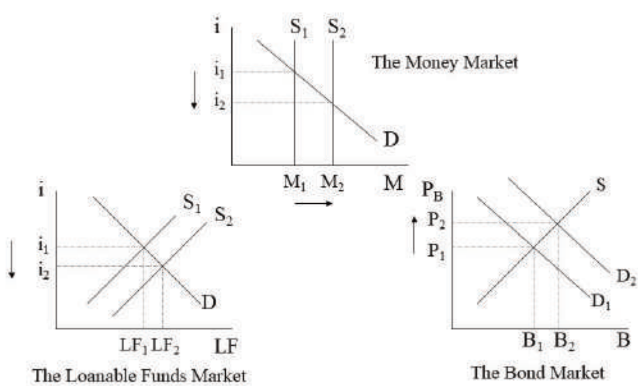
Another operation of commercial banks that has similar impacts on the M1 and M2 money supplies is the purchase and sale of government securities. In this case, government securities act in the role of non-money assets like promissory notes. Otherwise, the process is essentially the same. Banks purchase government securities from bond dealers and credit their checking accounts to pay them, which expands the M1 and M2 money supplies. When banks sell government securities to bond dealers, the dealers pay using their checkable deposits, and the M1 and M2 money supplies contract. Non-money assets return to circulation, and money assets are taken out of circulation.

We can now see exactly how the financial markets are affected due to the different operations we have been discussing. The previous discussion has shown that the M1 money supply can be increased due to the following bank operations:

1. A cash withdrawal from a savings account (or other non-M1 M2 asset account)
2. A transfer from a savings account (or other non-M1 M2 asset account) to a checking account
3. A loan granted to a borrower
4. A purchase of government securities from a bond dealer who is paid with cash or with a check that is then deposited in a checking account

If any of these transactions takes place, the resulting increase in the M1 money supply will have a direct impact on the financial markets. Figure 16.14 shows how the monetary expansion connects with our theory of financial markets developed in Chapter 15.

Figure 16.14: The Impact of a Monetary Expansion on the Financial Markets



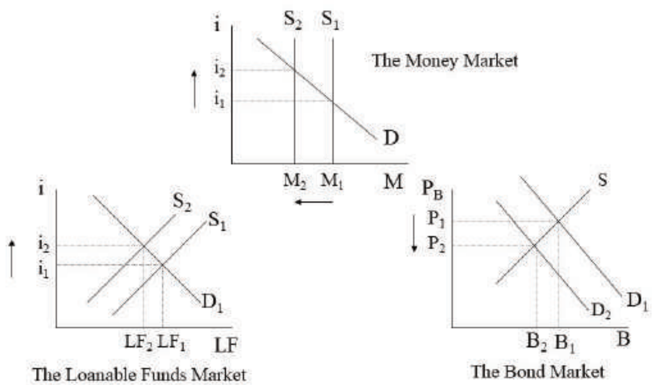
As Figure 16.14 shows, when the M1 money supply rises, the money supply curve shifts rightward in the money market. At the initial interest rate, a surplus of money exists. Because holders of money hold excess money balances, they decide to lend it in the loanable funds market to earn interest. The rightward shift of the supply curve for loanable funds creates a surplus of loanable funds. Competition between borrowers and lenders in the loanable funds market drives down the rate of interest, which clears the money market. At the same time, the rise in the supply of loanable funds is equivalent to an increase in the demand for bonds since these two markets represent a mirror reflection of one another. The higher demand for bonds pushes bond prices upward. Overall, the monetary expansion leads to an expansion of the loanable funds market and the bond market. Interest rates drop, and bond prices rise.

Similarly, the M1 money supply will decline due to the opposite bank operations:

1. A cash deposit into a savings account (or other non-M1 M2 asset account)
2. A deposit of a check drawn against a checking account into a savings account (or other non-M1 M2 asset account)
3. A transfer from a checking account into a savings account (or other non-M1 M2 asset account)
4. A loan repayment
5. A sale of securities to a bond dealer who pays with cash or check drawn against a checking account

If any of these transactions takes place, the resulting decrease in the M1 money supply will have a direct impact on the financial markets. Figure 16.15 shows how the monetary contraction connects with our theory of financial markets developed in Chapter 15.

Figure 16.15: The Impact of a Monetary Contraction on the Financial Markets



As Figure 16.15 shows, when the M1 money supply falls, the money supply curve shifts leftward in the money market. At the initial interest rate, a shortage of money exists. Because holders of money experience a shortage of money balances, they decide to lend fewer funds in the loanable funds market. The leftward shift of the supply curve for loanable funds creates a shortage of loanable funds. Competition between borrowers and lenders in the loanable funds market drives up the rate of interest, which clears the money market. At the same time, the reduction in the supply of loanable funds is equivalent to a reduction in the demand for bonds since these two markets represent a mirror reflection of one another. The drop in the demand for bonds pushes bond prices downward. Overall, the monetary contraction leads to a contraction of the loanable funds market and the bond market. Interest rates rise, and bond prices fall.[20]

One other aspect of the neoclassical theory of banking that is important is the multiple expansion of bank deposits through lending. That is, when First Regional Bank obtains new reserves and makes a loan, the new checkable deposit that is created adds to the M1 money supply. The process does not stop there, however, because the recipient of the check that is written against those funds (i.e., whomever the borrower of the funds decides to pay to obtain commodities) is likely to deposit the check in a Second Regional Bank checking account. Second Regional Bank will acquire new reserves. If Second Regional Bank lends the excess reserves, then another new checkable deposit is created. A check is then written against those funds and the recipient of the check might deposit the check in Third Regional Bank. Third Regional Bank thus receives new reserves and might lend the excess reserves, thereby creating yet another new checkable deposit. This process continues and each time a new checkable deposit (i.e., new money) is created. This process is referred to as the **multiple expansion of bank deposits**.

To understand the exact quantitative relationship between new reserves in the banking system and the amount of new checkable deposits or new money that is created, we need to look at the impact of these successive rounds of lending on the banks' balance sheets. Consider the changes to First Regional Bank's balance sheet when the bank accepts a $2,000 cash deposit and obtains $2,000 in new reserves. Assume that it is also subject to a

10% reserve requirement and so grants a $1,800 loan using the new excess reserves and that the borrower withdraws the entire amount from his checking account once the loan is granted. All these changes are depicted in Figure 16.16.

Figure 16.16: First Regional Bank Accepts a Deposit and Grants a Loan

Assets	Liabilities + Bank Capital
New Reserves +$2,000	Checkable Deposits +$2,000
Loan Asset +$1,800	Checkable Deposits +$1,800
Reserves -$1,800	Checkable Deposits -$1,800

In Figure 16.16, the deposit is made, and the bank acquires new reserves. It then grants a loan, acquiring a loan asset and creating a checkable deposit. Finally, the borrowed amount is withdrawn from the checking account, and the reserves decline by an equal amount. The reader should verify that the net effect is a $2,000 increase in bank assets and a $2,000 increase in liabilities. The M1 money supply has risen by $1,800 due to the granting of the loan.

Now consider what happens to Second Regional Bank's balance sheet when the $1,800 loan amount is deposited there. Second Regional Bank acquires new reserves due to this deposit. Assume that it lends the excess reserves. If the reserve requirement is still 10%, then Second Regional Bank must keep $180 in the form of reserves. It then lends $1,620 to a worthy borrower. The loan takes the form of a new checkable deposit which is then withdrawn. Figure 16.17 shows how these changes affect the balance sheet of Second Regional Bank.

Figure 16.17: Second Regional Bank Accepts a Deposit and Grants a Loan

Assets	Liabilities + Bank Capital
New Reserves +$1,800	Checkable Deposits +$1,800
Loan Asset +$1,620	Checkable Deposits +$1,620
Reserves -$1,620	Checkable Deposits -$1,620

The M1 money supply rises by $1,620 due to Second Regional Bank granting the loan and creating the new deposit. Let's finally consider how Third Regional Bank's balance sheet is affected if the $1,620 of borrowed funds is deposited there. Again, Third Regional Bank will acquire new reserves equal to $1,620. It will lend 90% of the newly acquired reserves (i.e., the excess reserves). The $162 of required reserves must be kept in the form of cash reserves. Second Regional Bank will lend $1,458 to a borrower and create a new checkable deposit, which adds to the M1 money supply. It will then be withdrawn, and the process continues. The changes to Third Regional Bank's balance sheet are shown in Figure 16.18.

Figure 16.18: Third Regional Bank Accepts a Deposit and Grants a Loan

Assets	Liabilities + Bank Capital
New Reserves +$1,620	Checkable Deposits +$1,620
Loan Asset +$1,458	Checkable Deposits +$1,458
Reserves -$1,458	Checkable Deposits -$1,458

We could explore the roles of Fourth Regional Bank, Fifth Regional Bank, and Sixth Regional Bank, and on and on, but the pattern should be clear by now.
$\Delta M1 = \$1,800 + \$1,620 + \$1,458 + \ldots$

The M1 money supply increases with each successive round of lending and deposit creation. With each round of lending, however, the new loan amounts become smaller and smaller. The reason for the reduction in deposit creation is that banks must hold a fraction of their newly acquired reserves each time they receive the deposited funds. The legal reserve requirement limits the banks to the creation of new money equal only to their excess reserves. As the story continues, the newly created deposits become smaller and smaller until they cease to influence the M1 money supply.

To arrive at an exact calculation of the total new checkable deposits that are created due to the acquisition of new reserves in the first round, we consider a situation in which First Regional Bank acquires only \$1.00 of new reserves. If R is the reserve requirement (10% in our example), then R (= 0.10) represents the required reserves in this case (i.e., 10 cents or \$0.10). The excess reserves that First Regional Bank acquires are 1-R (= 90 cents or \$0.90). The loan amount or newly created checkable deposit is assumed to equal the full amount of the excess reserves and so is also equal to 1-R. The first row of Table 16.2 shows the results for First Regional Bank.

Table 16.2: The Impact of \$1 of New Bank Reserves on the M1 Money Supply

Bank	New Reserves	Required Reserves (R)	Excess Reserves	New Checkable Deposits
First Regional	\$1.00	R	$1-R$	$1-R$
Second Regional	$1-R$	$R(1-R)$	$(1-R)^2$	$(1-R)^2$
Third Regional	$(1-R)^2$	$R(1-R)^2$	$(1-R)^3$	$(1-R)^3$
Fourth Regional	$(1-R)^3$	$R(1-R)^3$	$(1-R)^4$	$(1-R)^4$
Others	...	...	...	...

Second Regional Bank then acquires the 1-R amount of borrowed funds, which for it constitute new reserves. Its new required reserves are R times this amount or R(1-R). The excess reserves are 1-R times this amount or (1-R)(1-R) = $(1\text{-}R)^2$. The excess reserves are then loaned to a borrower and a new deposit is created in the amount of $(1\text{-}R)^2$. The second row in Table 16.2 captures these results.

Third Regional Bank then acquires the borrowed amount of $(1\text{-}R)^2$ in the form of reserves. Its required reserves are R times this amount or $R(1\text{-}R)^2$ and its excess reserves are 1-R times the new reserves or $(1\text{-}R)(1\text{-}R)^2 = (1\text{-}R)^3$, which is then loaned to a borrower. The third row in Table 16.2 captures these results. The results for Fourth Regional Bank have also been added to the table. Readers should think through those results to test their understanding.

To determine the total amount of new deposits created, we need to add up all the entries in the last column, which represents new money created. The change in deposits (D) can be represented as follows:

$$\Delta D = (1-R) + (1-R)^2 + (1-R)^3 + (1-R)^4 + ...$$

We can then factor out (1-R), which makes it possible to solve for D as follows:

$$\Delta D = (1-R)\{1 + (1-R) + (1-R)^2 + (1-R)^3 + (1-R)^4 + ...\}$$

$$\Delta D = (1-R)(1 + \Delta D)$$

$$\Delta D = (1-R) + (1-R)\Delta D$$

$$\Delta D\{1 - (1-R)\} = (1-R)$$

$$\Delta D = \frac{1-R}{1-(1-R)} = \frac{1-R}{R} = (1-R) \cdot \frac{1}{R} = \Delta E \cdot \frac{1}{R}$$

The change in deposits are equal to the change in excess reserves (E) times the **simple deposit multiplier** (1/R), which is also called the **money multiplier**.

This important result allows us to obtain an exact quantitative solution to the problem of the amount of deposits that can be created using the newly acquired reserves at the First Regional Bank. In that example, the reserve requirement is 10% and so the simple deposit multiplier is simply the reciprocal of the reserve requirement or 1/0.10 = 10. The newly acquired excess reserves at the First Regional Bank are equal to $1,800. Although this bank can only create $1,800 of new money using these excess reserves, the entire banking system can create ten times this amount, or $18,000. The M1 money supply thus rises by $18,000 once the money multiplier process has run its course. In general, when a loan is granted, the M1 money supply curve shifts rightward in the money market, but it may shift by an amount that is much larger than the initial loan amount due to the multiple expansion of bank deposits.

Two factors may weaken the money multiplier process. First, banks frequently choose to hold excess reserves to maintain liquid positions. Because these reserves exceed the minimum legal requirement, they reduce the impact of the money multiplier. If banks hold more reserves, then fewer funds are loaned out, and fewer new deposits are created. Second, if borrowers choose to hold some of their borrowed funds in the form of currency (or the recipients of those borrowed funds when the borrowers pay them for commodities), then the money multiplier process weakens. Currency that is held and not redeposited in the banking system does not add to the reserves of banks and cannot be loaned or used to create new deposits. Because the tendencies of banks to hold excess reserves or the public to hold currency are very real, the money multiplier that we developed in this section is almost certainly larger than we would observe. Advanced money and banking textbooks develop a more complex money multiplier that considers the existence of excess reserves and the currency leakages that occur when the borrowing public decides to hold cash.

Clearly the activities of banks and depositors have a major influence on the M1 money supply. At the same time, neither banks nor the depositors act with the intention of influencing the M1 money supply. According to neoclassical economics, banks are profit-seeking organizations and make decisions about loans and purchases and sales of securities without regard to the impact on the total stock of money in the economy. Similarly, depositors are utility-maximizing consumers and savers and make their decisions about the amount of money to hold without any concern for how their decisions might alter the money supply. Since the state of the economy can sometimes causes banks and depositors to act in predictable ways, the M1 money supply might be subject to extreme fluctuations. For example, during an economic downturn, banks restrict their lending, which reduces the M1 money supply and may intensify a recession. During an economic upswing, banks expand their lending, which might sharply increase the M1 money supply and may cause inflation. Similarly, if depositors decide to hold more currency during a recession, then banks will have fewer reserves to lend, which can worsen a recession. If they decide to deposit more currency in savings deposits (and other non-M1 M2 assets) during an economic expansion, then banks will have more reserves and are likely to grant more loans, which will expand the M1 money supply overall through the multiple expansion of checkable deposits. The potential impact on the M1 money supply stemming from the activities of banks and the public raises the question as to whether the determination of M1 should be left to the private banking community. The history of bank panics and bank runs in the United States provides part of the justification for the establishment of a central bank to regulate the money supply. The role of the central bank in money supply regulation is the subject of Chapter 17.

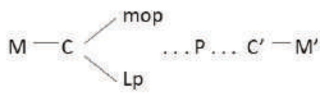
A Marxist Theory of Commercial Banks and its Theory of Financial Markets[21]

The major difference between the neoclassical and Marxist theories of commercial banking and financial markets that we explore in this chapter is that Marxists assign a central role to class conflict in their analyses of commercial bank behavior and financial markets. The Marxist theory developed in this section is not one that Marx created, but the concepts that Marx shaped directly inspired it. We begin with a reminder about the general formula for capital that is discussed in Chapter 4. Symbolically, capital takes the form of money and commodities but becomes capital when it participates in the following movement:

$M - C - M'$

According to this symbolic representation, M' is equal to M+ΔM where M represents surplus value. The capitalist thus transforms a sum of money capital (M) into more money (M'). This circuit of money capital may be expanded, as described in Chapter 4, to show that the money capital is advanced for the purchase of labor-power (Lp) and means of production (mop). Figure 16.19 shows the expanded circuit of money capital.

Figure 16.19: The Formula for the Circuit of Money Capital



After circulation is interrupted with the production phase of the process (P) in Figure 16.19, the finished commodity (C') is then transformed into a larger sum of money capital (M'). The surplus value is thus created in production and realized in exchange after production is complete. The reader should recall that it is the exploitation of labor-power that makes surplus value production and realization possible.

Industrial capitalists are capitalists who hire workers to produce commodities containing surplus value. Other kinds of capitalist exist as well. These capitalists receive a share of the surplus value that industrial capitalists appropriate in exchange for supporting the capitalist production process. For example, **commercial capitalists** (or merchant capitalists) hire workers who do not produce new commodities but who help to market and sell commodities. Because they assist with the realization of surplus value, industrial capitalists are willing to provide them with a share of the surplus value created in production, which becomes **commercial profit**. **Moneylending capitalists** provide industrial capitalists with the money capital they require to purchase means of production and labor-power. Because they help make possible the production of commodities containing surplus value, industrial capitalists are willing to provide them with a share of the surplus value created in production, which becomes interest.

Marxists argue that these two forms of capital, **merchant capital** and **interest-bearing capital**, are the oldest forms of capital. Ironically, it is necessary to understand how industrial capital produces surplus value if one is to understand the place of merchant capital and interest-bearing capital in modern capitalist societies. To understand the reason for this ironic result, consider the formula for the circuit of interest-bearing capital:

$M - M - C - M' - M'$

This circuit involves the transfer of interest-bearing capital to an industrial capitalist in monetary form. This part of the circuit (M-M) constitutes the granting of a loan. The circuit of industrial capital then follows with commodities (C) purchased and sold for a larger sum of money capital (M'). The initial loan amount (M) is then returned to the moneylending capitalist along with interest (ΔM) since $M' = M + \Delta M$ as before. It is important to note that only the interest payment is shown here, as opposed to the entire surplus value produced, because it is the circuit of interest-bearing capital that is under investigation. In fact, the surplus value is a larger sum and the interest only represents a portion of the surplus value. That interest only represents a fraction of the surplus value (or profit) is a claim that was made in Chapter 15. Competition in the loan capital market determines the specific share that is paid as interest.

When we introduce commercial banks into the analysis, an additional type of capital comes into play. **Bank capitalists** advance bank capital with the aim of grabbing a share of the surplus value produced in the industrial sector. The circulation of bank capital is represented in Figure 16.20.

Figure 16.20: The Circulation of Bank Capital

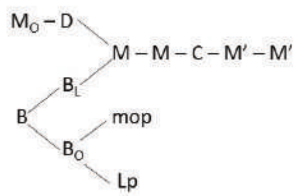


Figure 16.20 shows that depositors deposit their cash holdings (M_O) in a commercial bank, which become deposit capital (D). The bank capitalist also advances bank capital (B), which then splits into two parts. Part of the bank capital is needed for the purchase of labor-power (Lp) and means of production (mop) to operate the bank. This part of the bank capital is referred to as **bank**

operating capital (B_O). The labor-power that is hired to operate the bank is **unproductive labor-power** in the sense that these workers do not produce surplus value. Instead, they help to make the production of surplus value possible. The means of production that are purchased are not used to produce commodities containing surplus value. The means of production that a bank purchases include office supplies, computers, and electricity. For simplicity, it is assumed that all the capital is circulating capital, which allows us to omit fixed capital such as a banking facility.

It is essential to recognize that the bank operating capital ends its circulation at this stage. Because the means of production and labor-power are not used to purchase valuable commodities, this capital does not return to the banking capitalist in this part of the circuit. Another part of the bank capital is advanced for lending, however, and may be referred to as **bank loan capital** (B_L). It is combined with the deposit capital (D) to form the interest-bearing capital (M) described previously. The interest-bearing capital then passes through its circuit and returns with interest. The money capital that returns to the banking capitalist (M') is equal to the principal amount of the loan (i.e., the bank loan capital plus the deposit capital advanced) made at the beginning of the circuit (B_L+D) plus the gross bank profit (ΔM).

$$M' = B_L + D + \Delta M$$

In this case, the gross bank profit must be sufficient to pay the operating expenses of the bank (B_O) and return a net bank profit (ΔB) that is equal to the average profit in the economy.

$$\Delta M = B_O + \Delta B$$

Substituting the gross bank profit into the calculation of M' yields the following result:

$$M' = B_L + D + B_O + \Delta B$$

In other words, because the bank operating capital does not complete its circuit in the lower portion of Figure 16.20, it must return with the gross bank profit. If insufficient profit is realized to cover expenses and generate the average profit, then capitalists will cease to invest in the banking sector.

To understand how this analysis of commercial bank activity relates to the Marxist theory of financial markets presented in Chapter 15, we will initially simplify the discussion with the assumption that all bank capital is bank loan capital (i.e., no bank operating capital exists). Given this assumption, the principal amount of the loan (M) is equal to B+D since the entire bank capital is granted as a loan. The gross bank profit (ΔM) in this case is equal to the net bank profit (ΔB) because the bank need not earn enough profit to cover operating expenses, which are non-existent. We can use equations to express what has been stated thus far.

$$M = B + D$$

$$\Delta M = \Delta B$$

In Chapter 15, we explored a Marxist theory of interest rates. If we use the concepts developed in Chapter 16 to write an expression for the rate of interest, then it will be possible to link the two analyses. The **rate of interest** (i) is simply the amount of interest received on a loan (M) expressed as a percentage of the initial loan amount (M) as shown below:

$$i = \frac{\Delta M}{M} = \frac{\Delta B}{B+D}$$

At this stage, we only have a definition of the interest rate, and we lack a Marxist theory of its determination. To move in the direction of a theory, we divide the numerator and the denominator in the expression by the bank capital (B), which produces the following result:

$$i = \frac{\Delta M}{M} = \frac{\Delta B}{B+D} = \frac{\frac{\Delta B}{B}}{1+\frac{D}{B}}$$

To transform this definition of the rate of interest into a theory of the rate of interest, we need to recognize the numerator (i.e., the ratio of the net bank profit to the total bank capital) as the rate of profit on bank capital. In Marxian economics, competition among capitalists in an environment of unrestricted capital movements leads to an equalization of the profit rate throughout the economy. This rate of profit is referred as the **general rate of profit** (p). Of course, differences among profit rates across industries and capitalists always exist, and so the general rate of profit should be thought of as an average and the level towards which profits rates are constantly moving, even as unexpected factors cause deviations from this level.

Let's assume that the general rate of profit (p) prevails in the banking sector. If bankers earn the general rate of profit, then we can substitute the general rate of profit for the profit rate expression in the numerator of our interest rate expression. Doing so provides us with an expression that we may refer to as the **general rate of interest** (i_g).

$$i_g = \frac{\Delta M}{M} = \frac{\Delta B}{B+D} = \frac{\frac{\Delta B}{B}}{1+\frac{D}{B}} = \frac{p}{1+\frac{D}{B}}$$

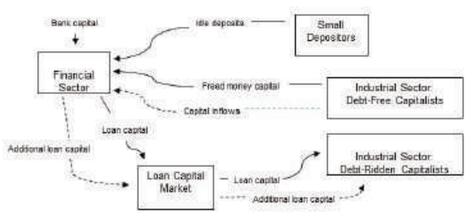
The general rate of interest is the rate of interest that banks must charge borrowers to ensure that they earn the general rate of profit. It is a hypothetical rate of interest rather than the observable, market rate of interest (i_m). The market rate of interest is what banks earn on their loans. If the market rate of interest exceeds the general rate of interest, then banks earn more than the general rate of profit. If the market rate of interest is below the general rate of interest, then banks earn less than the general rate of profit.

The factors that determine the general rate of interest can be easily gleaned from its expression. If the general rate of profit (found in the numerator in the expression) rises, then the general rate of interest must rise to ensure that sufficient interest is received to guarantee that the bank earns the higher profit rate. If the general rate of profit falls, then the general rate of interest must fall because it takes less interest to ensure that the bank earns the lower profit rate. In Chapter 14 we learned that Marxian economists argue that the long-term tendency of the general rate of profit to fall in capitalist economies causes economic crises. Although various factors counteract this long-term tendency, it is a tendency that reasserts itself periodically. If the general rate of profit has a long-term tendency to fall and the general rate of interest is directly related to the general rate of profit, then the general rate of interest also has a long-term tendency to fall, subject of course to various counteracting tendencies.

One other factor that influences the general rate of interest is the deposit/bank capital ratio (D/B), which is found in the denominator of the expression. If the deposit/bank capital ratio rises, then the general rate of interest falls. This result is intuitive. A higher deposit/ bank capital ratio means that banks are relying relatively more heavily on deposit capital when granting loans. These assets are very low-cost liabilities, particularly if they are checkable deposit liabilities that pay no interest, and so the banks require less interest to ensure that they receive the general rate of profit on bank capital. On the other hand, if the deposit/bank capital ratio falls, then banks are relying relatively more heavily on bank capital when granting loans and so must receive a higher rate of interest to ensure that the general rate of profit on bank capital is received.

It is one thing to assert that the general rate of interest is the rate of interest that will allow banks to earn the general rate of profit. It is another thing to assert that the market rate of interest will tend to move towards this level. To understand why the market rate of interest will gravitate towards the general rate of interest, consider what would happen if the market rate of interest was greater than the general rate of interest ($i_m > i_g$). Figure 16.21 illustrates how capitalists in the industrial sector and in the financial sector will respond to this discrepancy.

Figure 16.21: The Dynamics of Profit Rate Equalization Under Normal Conditions (market interest rate > general interest rate)



As previously explained, in this situation, banks will earn a profit rate that exceeds the general rate of profit. With unrestricted capital flows, industrial capitalists will begin to transfer capital out of industry and into finance. The inflow of capital into finance will cause banks to lend more in the loan capital market. The increased supply of loan capital will push down the market rate of interest in the direction of the general rate of interest. The process will continue until the market rate of interest equals the general rate of interest at which point the inflow of capital into finance will cease because the bank rate of profit is equal to the industrial rate of profit.

If the market rate of interest is below the general rate of interest, then the process works in reverse. That is, the bank rate of profit will be below the industrial rate of profit. Capital will flow out of finance and into industry. As the outflow of capital from finance continues, the supply of loan capital in the loan capital market will shrink, which puts upward pressure on the market rate of interest. The outflow of capital from finance will continue until the market rate of interest equals the general rate of interest at which point the bank rate of profit will once again equal the industrial rate of profit.[22]

It is also possible to incorporate the reserve requirement (R) into the analysis. If banks are required to keep a fraction of their deposit liabilities (D) in the form of reserves, then they will only be able to lend their excess reserves, which are equal to 1-R times the deposit liabilities. The expression for the general rate of interest, considering the reserve requirement, thus changes to the following:

$$i_g = \frac{\Delta M}{M} = \frac{\Delta B}{B+(1-R)D} = \frac{\frac{\Delta B}{B}}{1+\frac{(1-R)D}{B}} = \frac{p}{1+\frac{(1-R)D}{B}}$$

Our earlier expression for the general rate of interest is a special case of this expression where the reserve requirement is zero (R=0). Previously, banks could advance the entire deposit capital (D). Now they can only advance the excess reserves of (1-R)D. Mathematically, it should be clear that if R rises, then the general rate of interest will rise. That is, if banks are required to hold more reserves and thus lend less, then they must earn a higher interest rate on their loans to ensure that they receive the general rate of profit. If R falls, then banks can lend more and the rate of interest they require to obtain the general rate of profit is lower.

It is also worth noting that the maximum value of the general rate of interest is the general rate of profit. The maximum value is reached when banks are required to hold 100% of their deposits as reserves (R=1). In that case, they only lend bank capital and so must receive an interest rate equal to the general profit rate if they are to earn that profit rate. The minimum value of the general rate of interest is the expression we discussed previously where the reserve requirement was equal to zero (R=0). It is the lower bound because the banks can lend all their deposit capital to borrowers. The upper and lower bounds of the general rate of interest may be expressed as follows:

$$\frac{p}{1+\frac{D}{B}} \leq i_g \leq p$$

A final modification that we can make to the general rate of interest expression involves the incorporation of bank operating capital (B_O), which has been assumed to equal zero thus far. If bank operating capital is positive, then the bank must receive it as part of gross bank profit (ΔM). Also, if bank operating capital is positive, then the bank cannot lend its entire bank capital but only a portion that we call the bank loan capital (B_L). Making these adjustments to the general rate of interest expression and dividing the numerator and denominator by the bank capital yields the following result:

$$i_g = \frac{\Delta M}{M} = \frac{B_O+\Delta B}{B_L+(1-R)D} = \frac{\frac{B_O}{B}+p}{\frac{B_L}{B}+\frac{(1-R)D}{B}}$$

This new expression for the general rate of interest shows that a rise in bank operating capital will raise the general rate of interest. This result is expected because higher operating costs mean that banks require a higher interest rate to cover these expenses and obtain the general rate of profit on bank capital. Also, because the total bank capital equals the sum of bank operating capital and bank loan capital, a rise in bank operating capital involves a reduction in bank loan capital, which also raises the general rate of interest. That is, when banks lend less, they require a higher rate of interest to obtain the same general rate of profit. Conversely, a reduction in bank operating capital will lower the general rate of interest because banks have lower expenses and can manage with a lower rate of interest. They also lend more and so can obtain the general rate of profit when they charge a lower interest rate.

When the market rate of interest and the general rate of interest are equal, we have seen that no capital movements will occur between industry and finance. The reason is that the profit rate in the banking sector is equal to the profit rate in the industrial sector, and so capitalists have no incentive to reallocate capital across sectors. The static situation described here can be illustrated with the numerical example in Table 16.3 that uses the final expression for the general rate of interest developed in this chapter.

Table 16.3: A Numerical Example of Profit Rate Equalization Across Sectors

Industrial Sector		**Financial Sector**	
Industrial Rate of Profit	25.00%	Financial Rate of Profit	25.00%
Non-Borrowed Capital ($)	305,000	Bank Operating Capital ($)	5,000
Borrowed Capital (Loan Capital) ($)	195,000	Bank Loan Capital ($)	35,000
Total Productive Capital ($)	500,000	Total Deposits ($)	200,000
Total Profit ($)	125,000	Reserve Requirement Ratio	20.00%
Total Interest Paid ($)	15,000	**Total Bank Capital ($)**	40,000
Total Annual Social Product ($)	625,000	**General Rate of Interest (Prime Rate)**	≈7.69%
		Loan Interest Received or Gross Bank Profit ($)	15,000
		Net Bank Profit ($)	10,000
	Aggregate Profits ($)	135,000	
	Total Social Capital ($)	540,000	
	General Rate of Profit (Gross)	25.00%	

The two-sector example shown in Table 16.3 treats all items except for the bold-faced items as given. The bold-faced items are calculated using the definitions in this chapter. For example, it is assumed that a 25% rate of profit prevails in the two sectors. Summing together the given bank operating capital and bank loan capital, we obtain the total bank capital. To obtain the borrowed capital in the industrial sector, we calculate the sum of the bank loan capital and the excess reserves, $B_L+(1\text{-}R)D$, using the given required reserve ratio (R). The sum of the borrowed capital and the given non-borrowed capital in the industrial sector equals the total productive capital. The total industrial profit is calculated as the product of the industrial profit rate and the total productive capital. The general rate of interest is calculated using the final formula for the general rate of interest that we developed. The total interest paid in the industrial sector is the product of the general rate of interest and the borrowed capital. The total annual social product is the sum of the total productive capital and the total industrial profit. The gross bank profit is the same as the interest that industrial capitalists pay to the banking capitalists. The net bank profit is the gross bank profit minus the bank operating capital. The aggregate profits are the sum of the net bank profits and the industrial profits. The total social capital is the sum of the total productive capital and the total bank capital. The gross general rate of profit is the ratio of the aggregate profits to the total social capital. The reader should carry out these calculations to test his or her understanding of the relationships. Table 16.4 provides formulas for each of the calculations.

Table 16.4: The Formulas for the Static Case of Profit Rate Equalization Across Sectors

Industrial Sector		Financial Sector	
Industrial Rate of Profit	p	Financial Rate of Profit	p
Non-Borrowed Capital ($)	NBK	Bank Operating Capital ($)	B_O
Borrowed Capital (Loan Capital) ($)	$B_L+(1-R)D$	Bank Loan Capital ($)	B_L
Total Productive Capital ($)	K	Total Deposits ($)	D
Total Profit ($)	pK	Reserve Requirement Ratio	R
Total Interest Paid ($)	igM	**Total Bank Capital ($)**	B
Total Annual Social Product ($)	$K+pK$	**General Rate of Interest (Prime Rate)**	ig
		Loan Interest Received or Gross Bank Profit ($)	$B_O+\Delta B$
		Net Bank Profit ($)	ΔB
	Aggregate Profits ($)	$pK+pB$	
	Total Social Capital ($)	$K+B$	
	General Rate of Profit (Gross)	p	

To summarize the analysis in this section, the exploitation of labor-power in the industrial sector generates the surplus value that is appropriated as profit in the industrial sector and then as interest in the banking sector. The banking sector is parasitic in that it does not produce surplus value but helps make the production of surplus value possible when it grants loans to the industrial capitalists who use the funds to purchase labor-power and the means of production. To ensure the continuation of such loans, the industrial capitalists allow the banking capitalists to share in the surplus value that they appropriate from the working class.

It is helpful to reflect on the relationship between this Marxian analysis of financial markets and the analysis of the rate of interest in Chapter 15. In Chapter 15, a theory is presented that shows how the market rate of interest fluctuates over time as the level of production changes. That is, the business cycle is assumed in that analysis, and the rate of interest fluctuates due to discrepancies between the supply and demand for loanable money capital. In Chapter 16, however, the business cycle is not responsible for changes in the market rate of interest. It is capital's search for the highest rate of profit that is responsible for adjustments of the market rate of interest to the general rate of interest. Even if we assume an abstract situation in which the level of production is stable and the business cycle is not operative, the search for a maximum profit rate still applies, and capital will flow between industry and finance accordingly. The general rate of interest then is the average rate of interest towards which the market rate of interest moves. Deviations from this level are due to discrepancies in the loanable money capital market over the course of the business cycle.

Following the Economic News[23]

Colleges and universities in the United States have been entering into agreements with large banks like Wells Fargo and PNC Financial Services Group. The banks advertise their services on campus to acquire new accounts and then pay millions of dollars in royalties to the schools in exchange. The banks also collect a variety of fees from students, including ATM fees, which may then be partially distributed to the institutions of higher learning with which they are associated. From a neoclassical perspective, the banks are actively trying to acquire new deposit liabilities from the large student population. Those liabilities help the banks to raise funds that can be used to purchase income-earning assets, especially loan assets and securities. Because these income-earning assets are so lucrative, they are willing to pay royalties to the colleges and universities. The banks also generate revenue when they collect ATM fees and overdraft fees from the students. As the banks acquire more deposits from students, they also obtain additional reserves, which they can use to expand the money supply through lending and securities purchases. From a Marxian perspective, the banks are eager to acquire new funds from students so that they can grant interest-bearing loans to the productive, industrial sector. The interest received from these loans is a payment from the surplus value produced in the industrial sector. This redistributed surplus value is then transferred to colleges and universities for making the loans and the appropriation of surplus value possible. At the same time, the students are paying fees, which represent payments out of wages that students or their parents earn. Because wages represent payments for labor-power in Marxian theory, the fees are a way for banks to appropriate some of the value that was distributed to workers in the industrial sector of the economy. That is, the fees allow banks to acquire some of the value produced that was not appropriated as surplus value but rather advanced as variable capital for the purchase of labor-power.

Summary of Key Points

1. The three functions of money in neoclassical theory are medium of exchange, unit of account, and store of value.
2. The double coincidence of wants problem exists in barter economies because commodity owners must be able to find other commodity owners who possess the commodities they want and who are willing to exchange those commodities for what they own.

3. Commodity money is an object with intrinsic value, convertible paper money is paper that is convertible into commodity money, and inconvertible paper money is paper that is not convertible into anything else.
4. The M1 money supply includes publicly held currency and checkable deposits; the M2 money supply includes M1 and savings deposits, small time deposits, MMDAs, and MMMFs; the M3 money supply includes M2 and large time deposits.
5. The M_a definition of the money supply is like the M2 definition of the money supply except that: 1) it includes time deposits and federal savings bonds at current redemption rates; 2) it includes the cash surrender values of life insurance policies; and 3) it subtracts the demand deposits of thrift institutions to avoid double counting.
6. In a 100% reserve banking system, the entire money supply is backed up with bank reserves whereas in a fractional reserve banking system, only a fraction of the money supply is backed up with bank reserves.
7. The balance sheet of a bank lists its assets, liabilities, and net worth, and the total assets must always equal the sum of the liabilities and net worth.
8. Securities, loans, and cash reserves are the most important assets for banks. Deposits and borrowings are the most important liabilities for banks.
9. Successful banks must manage their assets,

liabilities, degree of liquidity, and bank capital simultaneously.

10. Banks are legally required to hold a specific fraction of their deposits as reserves, which is referred to as the reserve requirement.
11. Commercial banks increase the M1 money supply when they grant loans to borrowers, purchase bonds from bond dealers, and allow transfers from savings accounts into checking accounts or currency. The opposite actions reduce the M1 money supply.
12. When banks acquire new reserves, the rounds of successive lending that are initiated increase the M1 money supply by a multiple of the initial increase in bank reserves.
13. Whereas industrial capitalists appropriate the surplus value that their workers produce, commercial capitalists, moneylending capitalists, and bank capitalists receive a share of the surplus value that industrial capitalists appropriate.
14. The general rate of interest is the rate of interest that prevails when the rate of profit in the financial sector is the same as the rate of profit in the industrial sector.

List of Key Terms

Medium of exchange

Barter economy

Double coincidence of wants problem

Unit of account

Store of value

Commodity money

Convertible paper money

Inconvertible paper money

Fiat money

Digital currencies

Mining

Money supply

Money stock

Liquidity

Currency

M1 money supply

Thrift institutions

Savings and loan associations

Credit unions

M2 money supply

Certificates of deposit (CDs)

Money market mutual funds (MMMFs)

Money market deposit accounts (MMDAs)

M3 money supply

Cash surrender values

Austrian definition of the money supply (M_a)

100% reserve banking system

Bank run

Fractional reserve banking system

Balance sheet

Assets

Liabilities

Net worth (equity)

Balance sheet equation

Insolvent

Securities

Loans

Default

Loan losses

Cash reserves

Deposits

Borrowings

Federal funds market

Federal funds rate

Discount loans

Discount rate

Bank capital

Reserve requirement

Excess reserves

Multiple expansion of bank deposits

Simple deposit multiplier

Money multiplier

Industrial capitalists

Commercial capitalists

Commercial profit

Moneylending capitalists

Merchant capital

Interest-bearing capital

Bank capitalists

Bank operating capital

Unproductive labor-power

Bank loan capital

Rate of interest (i)

General rate of profit (p)

General rate of interest (i_g)

Problems for Review

1.Can you think of an asset that is a good store of value but not a good unit of account or medium of exchange? Can you think of an asset that is a good unit of account but not a good medium of exchange or store of value? Can you think of an asset that is a good medium of exchange but not a good unit of account or store of value?

2. Suppose that you are presented with the following information:

- Savings deposits = $9 trillion
- Cash surrender values of life insurance policies = $2.5 trillion
- Checkable deposits = $3 trillion
- Large time deposits (face value) = $1.5 trillion
- Small time deposits (face value) = $4 trillion
- Large time deposits (current redemption value) = $1 trillion
- Publicly held currency = $1.5 trillion
- Savings bonds (current redemption value) = $3 trillion
- MMDAs = $3.5 trillion
- MMMFs = $2.5 trillion
- Small time deposits (current redemption value) = $3.5 trillion

- Demand deposits that thrift institutions hold in commercial banks = $1 trillion

Calculate the M1 money supply.
Calculate the M2 money supply.
Calculate the M3 money supply.
Calculate the M_a money supply.

3. Suppose that you are presented with the following information:

- A bank has $25 billion in cash reserves.
- A bank has issued checkable deposits of $41 billion.
- A bank has borrowings of $62 billion.
- A bank has issued savings deposits of $42 billion.
- A bank has $48 billion in securities.
- A bank has $66 billion in loan assets.

Create a balance sheet for the bank. Determine the level of bank capital. Does the balance sheet equation hold with the answer you have provided? What is the bank's situation like?

4. Suppose that a depositor decides to transfer $3000 from a checking account to a savings account. Create a balance sheet for the bank, and show how the balance sheet will be affected due to this change. Only list the changes on your balance sheet. Then explain what will happen to the M1 and M2 money supplies.

5. Suppose a bank purchases $60,000 in bonds from a bond dealer who has a checking account at the bank. Create a balance sheet for the bank, and show how the

balance sheet will be affected due to this change. Only list the changes on your balance sheet. Then explain what will happen to the M1 and M2 money supplies.

6. Suppose that a borrower repays a loan in the amount of $80,000. Create a balance sheet for the bank, and show how the balance sheet will be affected due to this change. Only list the changes on your balance sheet. Then explain what will happen to the M1 and M2 money supplies.

7. Suppose that a borrower makes a cash withdrawal of $4000 from a savings account. Create a balance sheet for the bank, and show how the balance sheet will be affected due to this change. Only list the changes on your balance sheet. Then explain what will happen to the M1 and M2 money supplies.

8. Suppose banks acquire $50 billion in new reserves, and the reserve requirement ratio is 6%. What will be the impact on the total deposits in the system, assuming all excess reserves are loaned to borrowers and the public redeposits all the borrowed funds in the banking system?

9. Suppose that you know the following information about the banking system. Use the information to complete the rest of the table.

Table 16.5: Review Problem #9

Industrial Sector		**Financial Sector**	
Industrial Rate of Profit	20.00%	Financial Rate of Profit	20.00%
Non-Borrowed Capital ($)	450,000	Bank Operating Capital ($)	8,000
Borrowed Capital (Loan Capital) ($)		Bank Loan Capital ($)	42,000
Total Productive Capital ($)		Total Deposits ($)	290,000
Total Profit ($)		Reserve Requirement Ratio	15.00%
Total Interest Paid ($)		**Total Bank Capital ($)**	
Total Annual Social Product ($)		**General Rate of Interest (Prime Rate)**	
		Loan Interest Received or Gross Bank Profit ($)	
		Net Bank Profit ($)	
	Aggregate Profits ($)		
	Total Social Capital ($)		
	General Rate of Profit (Gross)		

Notes

1. The three functions of money discussed in this section are universally recognized in neoclassical textbooks.
2. Mishkin (2006), p. 46, emphasizes the reduction in search time as an efficiency-enhancing characteristic of money.
3. Mishkin (2006), p. 46, also refers to the promotion of specialization as a second efficiency-enhancing characteristic of money.
4. See Marx (1990), p. 227-244, for a detailed account of these functions.
5. This example is found in OpenStax (2014), p. 617.
6. The Federal Reserve concluded that “M3 does not appear to convey any additional information about economic activity that is not already embodied in M2” (Federal Reserve Statistical

Release. H.6 Money Stock Measures: Discontinuance of M3. November 10, 2005. Revised March 9, 2006. Web. Accessed on January 20, 2018. https://www.federalreserve.gov/Releases/h6/discm3.htm).

7. Rothbard (1978), p. 151.
8. Ibid. pp. 145.
9. Ibid. pp. 146-148.
10. Ibid. pp. 148.
11. Ibid. pp. 149.
12. Ibid. pp. 150.
13. Ibid. pp. 150.
14. Ibid. pp. 151.
15. Ibid. pp. 151.
16. Ibid. pp. 151-152.
17. This imaginative approach is inspired by the goldsmith story that is found in neoclassical textbooks. See Even, Louis. "The Goldsmith Who Became a Banker - A True Story." *Cahiers du Crédit Social. October 1936.*
18. It is common for neoclassical textbooks to present a consolidated balance sheet for commercial banks and then discuss the different balance sheet items. For example, see Samuelson and Nordhaus (2001), p. 522, and Mishkin (2006), p. 201-205. Mishkin (2006), p. 201-205, and Hubbard and O'Brien (2019), p. 872-874, provide detailed discussions of each balance sheet item, which are like the approach of this section.
19. Frederic Mishkin (2006), p. 208, identifies the corresponding types of management for these objectives as asset management, liability management, liquidity management, and capital adequacy management.
20. A higher demand for loanable funds and a higher supply of

bonds is also expected, but we emphasize the other shifts because their larger impact brings about contractions of the financial markets consistent with the monetary contraction.

21. The theory presented in this section draws from the following source: Saros, Daniel E. "The Circulation of Bank Capital and the General Rate of Interest." *Review of Radical Political Economics*, 45 (2), Spring 2013: 149-161. The final, definitive version is available at http://online.sagepub.com/. https://journals.sagepub.com/doi/full/10.1177/0486613412458647
22. The analysis presented here is a simplification. A complicating factor is that when capital flows into or out of finance, the level of bank capital changes, which alters the general rate of interest as well. I have explained how to analyze the simultaneous movements in the general rate of interest and the market rate of interest in Saros (2013), p. 158-159.
23. Korn, Melissa and Christina Rexrode. "Big Banks Pay Big Bucks for Top Billing on College Campuses." *The Wall Street Journal*. Web. Updated January 28, 2018. Accessed on February 3, 2018.

Chapter 17

MONETARY THEORIES AND THE ROLE OF THE CENTRAL BANK

Goals and Objectives:

In this chapter, we will do the following:

1. *Identify* the organizational structure and functions of the Federal Reserve
2. *Analyze* the tools that the central bank uses to influence the financial markets
3. *Examine* the neoclassical approach to monetary policy using an exogenous money supply
4. *Inspect* the connection between the Quantity Theory of Money and the AD/AS Model
5. *Explore* a Post-Keynesian approach to monetary policy using an endogenous money supply
6. *Investigate* a Marxian theory of fiat money and relate it to U.S. economic history
7. *Link* the monetary policy tools to the Marxian theory of financial markets

In Chapter 16, we explored different methods of measuring the quantity of money and how commercial banks influence those different measures. We also looked at how commercial banking activities influence the financial markets from neoclassical and Marxian perspectives. In this chapter, a major purpose is to explore the role of the central bank in the determination of the money supply. All modern capitalist economies have a central bank and so an analysis of modern banking systems must incorporate its role into any meaningful analysis. Once we understand the role the central bank plays in the determination of the money supply, we will then be able to understand how the central bank influences the overall economy. Of course, our understanding depends on our theoretical lens, and so we will explore this question from neoclassical, Post-Keynesian, and Marxian perspectives.

The Organizational Structure and Functions of the Federal Reserve

Modern capitalist nations have central banks that are responsible for the management of the money supply and the conduct of monetary policy. **Monetary policy** refers to the use of money supply changes to influence aggregate production, the aggregate price level, and the level of unemployment. In Japan, the central bank is the Bank of Japan. In England, the central bank is the Bank of England. In the eurozone (i.e., in the European nations that have adopted the euro), the central bank is the European Central Bank. In the United States, the central bank is the **Federal Reserve** (also known as the Fed).

The Federal Reserve was created by an Act of Congress in 1913. After the 1907 financial crisis, private bankers recognized that some central control of the money supply was needed to prevent bank panics from becoming major financial crises. The notion of a central bank that could intervene during a bank panic as a lender of last resort to stabilize the financial markets was perceived to be necessary. Rather than centralize financial power in a single central bank, however, a compromise was struck that created 12 **Federal Reserve Banks** in 12 distinct Federal Reserve districts throughout the nation. The 12 Federal Reserve Banks are in the following cities:

1. Atlanta
2. Boston
3. Chicago
4. Cleveland
5. Dallas
6. Kansas City
7. Minneapolis
8. New York
9. Philadelphia
10. Richmond
11. San Francisco
12. St. Louis

Each Federal Reserve Bank has a Board of Directors and a Board-appointed Federal Reserve Bank President. In addition to the 12 Federal Reserve Banks, the Federal Reserve System has its headquarters in Washington, DC. A **Federal Reserve Board of Governors**, which consists of appointed officials, governs the entire system. The Board of Governors consists of seven members that the President of the United States appoints for 14-year staggered terms subject to U.S. Senate confirmation. The staggering of the terms guarantees that one Governor's term expires every two years, which ensures a significant degree of continuity on the Board. If, for example, the terms of all seven Governors expired in the same year, then an entirely new Board would be appointed that year, and it would be difficult for the new Governors to benefit from the knowledge and experience of past Governors. The President of the United States also appoints one Governor to serve as **Chair of the Federal Reserve Board** and another Governor to serve as **Vice-Chair**. The terms of Chair and Vice Chair are four-year, renewable terms. Generally, when the term of a Chair expires, they are either reappointed or they leave the Board altogether. The position of Federal Reserve Board Chair is the most powerful position in the Federal Reserve system. The Fed Chair serves as the spokesperson for the Fed. Past Chairs have included Alan Greenspan, Ben Bernanke, and Janet Yellen. The Fed Chair in 2018 is Jerome Powell.

A second governing body within the Federal Reserve system is the **Federal Open Market Committee (FOMC)**. The FOMC consists of 12 members, including the seven members of the Board of Governors, the President of the New York Federal Reserve Bank, and four other Federal Reserve Bank Presidents who serve on a one-year, rotating basis. The FOMC meets about every 6 weeks at the Federal Reserve headquarters in Washington, DC and is responsible for conducting all Fed interventions in the bond market. Federal Reserve bond market interventions are conducted through the New York Fed, which is the reason that the New York Fed President is the only permanent member of the FOMC. The New York Fed is important for several other reasons, as Frederic Mishkin explains, including its role as a gold repository, its conduct of foreign exchange market interventions, and its proximity to the largest financial institutions in the world.[1]

The legal status of the Fed is also somewhat unusual. It is regarded as a **quasi-public institution** in that it is privately owned but publicly controlled. The owners of the Fed are the private commercial banks within the Federal Reserve system. The banks own stock in the Fed and may receive profit distributions each year if the Fed earns a profit, subject to maximum limits. The legal mandate of the Fed is not to make a profit, however, but to serve the public interest.[2] Control of the system rests with public officials, and so it is neither a purely private nor a purely public institution.

Because the officials that control the Federal Reserve system are appointees with long terms in office, the Fed operates with a considerable degree of independence. That is, elected officials have relatively little control of Fed policy. Of course, Congress can change the Federal Reserve Act or even abolish the Fed if it repeals that legislation, but such changes require major legislative action. As the law stands, elected officials can do little to alter monetary policy aside from waiting until new Governors are appointed to the Federal Reserve Board. The notion of **central bank independence** is a controversial one. Its strength is simultaneously its weakness. Independence allows the Fed to pursue politically unpopular policies that serve the public interest, but the lack of democratic control of these decision makers makes many people uncomfortable as it seems to conflict with the principles of American democracy. It does not seem possible to resolve this question to the satisfaction of all parties. In any case, the general trend throughout the world in recent decades has been towards greater central bank independence to avoid harmful economic policies, especially those that encourage inflation.

The Fed performs several key functions. Almost all Fed functions can be inferred from the elements on its balance sheet. The one function of the Fed that is not clearly reflected in the Fed's balance sheet is its role as a **supervisor of banks**. The Fed shares this responsibility with other state and federal government agencies, which involves monitoring levels of bank capital, reserves, and risk. Figure 17.1 shows the balance sheet of the Federal Reserve. The figure combines all assets, liabilities, and net worth for all 12 Federal Reserve Banks.

Figure 17.1: The Consolidated Statement of Condition of All Federal Reserve Banks (in millions of dollars)

Assets	Liabilities + Bank Capital
Coin $1,945	Federal Reserve Notes (outstanding) $1,560,302
Securities $4,201,346	Reserves $2,142,376
Loans $54	Treasury Deposits 275,794
Foreign currency-denominated assets $22,110	Other $399,266
Items in process of collection $91	Bank capital $41,487
Other $193,679	
TOTAL = $4,419,225	TOTAL = 4,419,225

Source: Board of Governors of the Federal Reserve System (U.S.), Federal Reserve Statistical Release. H.4.1 Factors Affecting Reserve Balances. Web. Release Date: 2/1/2018. Retrieved on 2/6/2018. https://www.federalreserve.gov/releases/h41/current/h41.htm#h41tab9

On the asset side of the Fed's balance sheet, we see loans to commercial banks. The loans to commercial banks represent the original function of the Fed as **lender of last resort**. During a financial panic, banks cease to lend due to fear that borrowers will default on those loans. By granting loans to banks, the Fed can help restore confidence in the financial system. Banks will feel reassured that the banks to whom they lend will receive assistance from the Fed, if necessary, so that they can repay their loans with interest. Because the loans that the Fed grants to banks are interest-bearing loans, they are assets for the Federal Reserve. The reader might be surprised that the loan assets of the Fed only amount to $54 million, which is a small sum in comparison with other items on the asset side of the balance sheet. Because the U.S. economy is not experiencing a financial crisis, the lender of last resort role of the Fed is not especially necessary. Another important reason for the low figure is that the Fed encourages banks to borrow from other banks in the federal funds market. It does so by charging an interest rate on loans to banks that is higher than the federal funds rate. Therefore, banks voluntarily choose to borrow from one another rather than from the Fed.

The Federal Reserve is also an **issuer of currency**. It is important to note that the Fed does not print money, which is a common misconception. The Fed does not print U.S. dollars. It is the Bureau of Engraving and Printing within the U.S. Department of Treasury that is responsible for that duty. Once the dollars are printed, however, a Federal Reserve Bank must issue them. If you look at a U.S. dollar, you can see which Federal Reserve Bank issued your bill. You will also see the words "Federal Reserve Note" at the top of your U.S. dollar. In fact, **Federal Reserve Notes** are U.S. dollars. The function of the Fed as an issuer of currency also shows up on the liabilities side of the Fed's balance sheet. The Federal Reserve Notes (outstanding) entry on the balance sheet shows how many U.S. dollars have been issued. It might seem strange that Federal Reserve Notes are treated as a liability of the Fed. If the Fed notes represented convertible paper money, then it would be easy to understand why they are treated as liability. They would represent a claim to so much gold or silver on the asset side of the Fed's balance sheet. Because U.S. dollars are fiat money, they do not represent a claim to gold, but from a conceptual point of view, you can think of them as representing a claim to the assets of the Fed even though they are not directly redeemable. Incidentally, the Fed does own gold and other precious metals, which are included in the Other category on the asset side of the Fed's balance sheet.

The reader might be puzzled that coins are found on the asset side of the Fed's balance sheet even as U.S. dollars are treated as a liability of the Fed. The reason is that while Federal Reserve Banks issue U.S. dollars, the Fed does not issue coins. The U.S. Treasury mints and issues coins. Therefore, coins are a liability of the Treasury, and they are treated as an asset when the Fed acquires them.

The Fed also serves as a **banker for commercial banks**. This role is reflected in the reserves of commercial banks, which is a liability on the Fed's balance sheet. The Fed owes banks that transfer currency and other assets to the Fed. For example, if the Fed buys securities from a commercial bank, then it may pay the bank by crediting the bank's reserve account. This operation increases the Fed's assets and liabilities. It reduces the bank's assets because it loses securities, but it also increases its assets by an equal amount because it gains reserves. Alternatively, the bank might transfer currency to the Fed, which leaves the Fed's liabilities unchanged. It reduces the Fed's liabilities because its outstanding Federal Reserve Notes decline, but its liabilities increase by an equal amount as it credits the bank's reserve account. The bank loses currency but gains reserves and so its assets remain the same overall. For many years, banks were not paid interest for the reserves held at the Fed. In 2008, this policy changed, and banks have since earned interest on their Fed reserve holdings.

The Fed also serves as a **processor of checks** as explained in Chapter 16. When an account holder writes a check against a checkable deposit, the bank that receives the check sends it to the Fed for processing. The Federal Reserve Bank in that district increases the reserves of the recipient bank and debits the reserves of the bank with the account against which the check is drawn. The changes to bank reserves do not occur immediately, however, and so the reserves of the recipient bank increase before the reserves of the other bank are debited. During this period, the Fed acquires a new asset referred to as items in process of collection. The items in process of collection on the Fed's balance sheet thus reflects this function of the Fed.

The Fed also serves as a **banker for the federal government**. Individuals and firms cannot open bank accounts at Federal Reserve Banks. Only commercial banks in the Federal Reserve system and the U.S. government have this privilege. The U.S. Treasury collects hundreds of billions of dollars in tax revenue each year. These funds must be deposited somewhere, and the U.S. Treasury account at the Fed is where these funds are deposited. U.S. Treasury deposits are a liability for the Fed because the Fed owes the Treasury once these deposits are received. If they are transfers from commercial banks, then the Fed's reserve liabilities decline and its Treasury deposit liabilities increase by an equal amount.

One other asset that appears on the Fed's balance sheet is **foreign currency-denominated assets**. Such assets reflect the Fed's history of intervention in foreign exchange markets. The Fed has the power to intervene on a massive scale in foreign currency markets. By doing so, it can alter the foreign exchange values of the U.S. dollar and other currencies for which it trades. A deliberate manipulation of the U.S. exchange rate can make U.S. exports (or U.S. imports) cheaper or more expensive. Such changes can have an impact on aggregate spending in the U.S. and thus aggregate output, employment, and the price level as discussed in Chapter 13. In an era of floating exchange rates, however, the Fed is not actively involved in managing the foreign exchange value of the U.S. dollar even though it retains the power to do so.

The Primary Central Bank Tools of Monetary Policy

The last function of the Fed that we can observe in the Fed's balance sheet is its function as **regulator of the money supply**. This regulatory function includes the three primary tools of monetary policy.[3] The first monetary policy tool and the tool that has been the most important in recent decades is open market operations. **Open market operations** refer to the Fed's bond market interventions (i.e., the purchases and sales of government securities). When the Fed purchases government bonds from commercial banks, it pays the banks with reserves (i.e., it credits their reserve accounts). The additional reserves can be used to grant loans, which expands the money supply. Alternatively, if the Fed sells government securities to commercial banks, then banks will pay for the bonds using reserves (i.e., it debits their reserve accounts). Faced with dwindling reserves, banks contract their lending, and the money supply falls. Hence, the securities and reserves items on the Fed's balance sheet represent the open market operations of the Fed and thus its role as regulator of the money supply.

A second monetary policy tool of the Fed is discount lending. **Discount loans** are loans that the Fed grants to commercial banks. Sometimes these loans are granted during financial crises and are consistent with the lender of last resort function. Sometimes the loans are granted during periods of economic and financial stability but to banks that are nevertheless in financial trouble. When the loans are granted, the Fed adds to the banks' reserve accounts. The banks that receive these reserves can then grant additional loans to borrowers, which expands the money supply. Although the Fed cannot force banks to borrow, the Fed can reduce the **discount rate**, which is the interest rate charged on discount loans. Because the loans are cheaper when the discount rate falls, banks are more inclined to borrow from the Fed. Alternatively, if the discount rate increases, then banks are less inclined to borrow from the Fed, and discount lending and bank lending contract, which reduces the money supply. Hence, the loans to commercial banks and the reserves of commercial banks on the Fed's balance sheet reflect the Fed's role as regulator of the money supply.

A third monetary policy tool of the Fed is the **reserve requirement ratio** (R). When the Fed lowers the reserve requirement ratio, commercial banks are not required to hold as many reserves. Commercial banks thus have more excess reserves, which they lend to borrowers, thereby expanding the money supply. Alternatively, if the Fed raises the reserve requirement ratio, then banks have fewer excess reserves, and they need to hold more reserves in their accounts with the Fed. To do so, they contract their lending, which reduces the money supply. Hence, the reserves of commercial banks on the Fed's balance sheet reflects the Fed's role as regulator of the money supply.

To see the impact of each of these monetary policy tools on the money supply, just consider the balance sheets of a commercial bank and the Fed. When the Fed engages in open market operations, its bond market purchases and sales influence the reserves in the banking system. For example, suppose that the Fed buys \$8,000 in bonds from commercial banks. The impact on the Fed's balance sheet and the commercial bank's balance sheet are shown in Figure 17.2.

Figure 17.2: The Fed Purchases Securities from a Bank

Figure 17.2 (a): The Fed gains securities and pays with reserves

Assets	Liabilities + Bank Capital
Securities +\$8,000	Reserves +\$8,000

Figure 17.2 (b): The bank loses securities and gains reserves

Assets	Liabilities + Bank Capital
Securities -\$8,000 Reserves +\$8,000	

Figure 17.2 shows that the Fed gains securities in the amount of \$8,000 and then credits the account of the bank as payment for the securities. The Fed's assets and

liabilities thus increase by an equal amount. The commercial bank, on the other hand, loses an asset when it sells the securities, but it gains an asset in the form of additional reserves. If the bank was meeting its reserve requirement previously, then these additional reserves constitute excess reserves for the bank. If the bank uses the excess reserves to grant loans, then new deposits are created and the M1 money supply expands. This process is also likely to lead to additional lending by other banks and the multiple expansion of bank deposits described in Chapter 16.

Now consider what happens when the Fed sells $4,000 in bonds to commercial banks. The impact on the Fed's balance sheet and the commercial bank's balance sheet are shown in Figure 17.3.

Figure 17.3: The Fed Sells Securities to a Bank

Figure 17.3 (a): The Fed loses securities and the bank pays using reserves

Assets	Liabilities + Bank Capital
Securities -$4,000	Reserves -$4,000

Figure 17.3 (b): The bank gains securities and loses reserves

Assets	Liabilities + Bank Capital
Securities +$4,000 Reserves -$4,000	

Figure 17.3 shows that the Fed loses securities in the amount of $4,000 and then debits the account of the bank

that is paying for the securities. The Fed's assets and liabilities thus decrease by an equal amount. The commercial bank, on the other hand, gains an asset when it buys the securities, but it loses an asset in the form of a reduction in reserves. If the bank was meeting its reserve requirement previously, then this loss of reserves constitutes a shortfall of reserves for the bank. The bank will be under pressure to contract its lending. As borrowers repay loans, it will not renew them. Its deposit liabilities will then decline, and the M1 money supply falls. This process is also likely to lead to a contraction of lending by other banks and a multiple contraction of bank deposits.

Now consider a situation in which the Fed reduces the discount rate that it charges banks for discount loans. If the discount rate is equal to the federal funds rate, then banks will have an incentive to borrow from the Fed when the discount rate is cut.[4] That is, it will be cheaper to borrow from the Fed than from other banks. Suppose that a reduction in the discount rate causes a bank to borrow $20,000 in discount loans from the Fed. Figure 17.4 shows the impact on the bank's balance sheet.

Figure 17.4: The Fed Grants a Discount Loan to a Bank

Figure 17.4 (a): The Fed grants a discount loan to a bank in the form of reserves

Assets	Liabilities + Bank Capital
Discount loan +$20,000	Reserves +$20,000

Figure 17.4 (b): The bank gains reserves, but the discount loan is a liability for the bank.

Assets	Liabilities + Bank Capital
Reserves +$20,000	Discount loan +$20,000

As Figure 17.4 shows, the Fed gains a loan asset and credits the bank's reserve account, which is a liability for the Fed. The bank, on the other hand, acquires a new asset in the form of additional reserves. Because the loan must be repaid, it is a liability for the bank. The additional reserves become excess reserves if the bank was meeting its reserve requirement previously. If it lends the excess reserves, then the lending that it encourages results in the multiple expansion of bank deposits and an increase in the M1 money supply.

Alternatively, suppose that the Fed increases the discount rate that it charges banks for discount loans. If the discount rate is initially equal to the federal funds rate, then an increase in the discount rate will lead to less borrowing from the Fed. That is, it will be cheaper to borrow from other banks than from the Fed. Suppose that an increase in the discount rate causes a bank to repay $30,000 in discount loans from the Fed. Figure 17.5 shows the impact on the bank's balance sheet.

Figure 17.5: A Bank Repays the Fed for a Discount Loan

Figure 17.5 (a): The Fed accepts the repayment of a discount loan in the form of reserves

Assets	Liabilities + Bank Capital
Discount loan -$30,000	Reserves -$30,000

Figure 17.5 (b): The bank repays a discount loan using reserves

Assets	Liabilities + Bank Capital
Reserves -$30,000	Discount loan -$30,000

As Figure 17.5 shows, the Fed will lose a discount loan asset, but its reserve liabilities will fall as the bank repays the loan using reserves. The bank, on the other hand, rids itself of a liability because it no longer owes the Fed for the discount loan, but its reserves fall on the asset side of its balance sheet because it uses reserves to repay the Fed for the loan. If the bank was just meeting its reserve requirement before, it now falls below that legal requirement. The bank will contract its lending as a corrective action. The contraction of bank lending for this one bank triggers a contraction of bank lending throughout the banking system and a reduction in checkable deposits. As a result, the M1 money supply contracts.

Finally, consider what happens when the Fed changes the reserve requirement ratio (R). Suppose that it reduces the reserve requirement from 10% to 8%. The impact on a bank's balance sheet is shown in Figure 17.6.

Figure 17.6: The Fed Reduces the Reserve Requirement

Figure 17.6 (a): A bank is just meeting the reserve requirement (R) of 10%

Assets	Liabilities + Bank Capital
Required Reserves $10,000 Excess Reserves $0 Total Reserves $10,000	Checkable deposits $100,000

Figure 17.6 (b): A bank possesses excess reserves when the Fed cuts R to 8%

Assets	Liabilities + Bank Capital
Required Reserves $8,000 Excess Reserves $2,000 Total Reserves = $10,000	Checkable deposits $100,000

As Figure 17.6 shows, the bank has acquired excess reserves of $2,000 due to the reduction in the reserve requirement. Originally, the bank has no excess reserves. As soon as the reserve requirement declines, however, the bank finds itself with excess reserves that it can lend. If it lends the excess reserves, then the bank's checkable deposits will increase. Therefore, the M1 money supply will rise.

Alternatively, consider what happens when the Fed raises the reserve requirement from 10% to 12%. The impact on a bank's balance sheet is shown in Figure 17.7.

Figure 17.7: The Fed Raises the Reserve Requirement

Figure 17.7 (a): A bank is just meeting the reserve requirement (R) of 10%

Assets	Liabilities + Bank Capital
Required Reserves $10,000 Excess Reserves $0 Total Reserves $10,000	Checkable deposits $100,000

Figure 17.7 (b): A bank falls below its required reserves when the Fed raises R to 12%

Assets	Liabilities + Bank Capital
Required Reserves $12,000 Excess Reserves -$2,000 Total Reserves $10,000	Checkable deposits $100,000

In the case of an increase in the reserve requirement, the bank finds itself with insufficient reserves. To increase the ratio of bank reserves to deposit liabilities, the bank has several options, but one option is to allow borrowers to repay loans while denying them new loans. As borrowers repay loans, the bank's checkable deposit liabilities will decline. If the process goes far enough, then the bank's reserves will provide sufficient backing for the checkable deposits that it has issued. The contraction of checkable deposits will lead to a contraction of lending throughout the banking system and to a multiple contraction of bank deposits. The result is a reduction in the M1 money supply.

We have now seen how the Fed's three monetary policy tools influence the M1 money supply. The Fed's monetary policy can be described as either expansionary or contractionary. If the Fed uses its policy tools to increase the money supply, then its monetary policy is described as an **expansionary monetary policy** or as an **easy monetary policy**. If, on the other hand, the Fed uses its policy tools to reduce the money supply, then its monetary policy is described as a **contractionary monetary policy** or as a **tight monetary policy**.

The Neoclassical Approach to Monetary Policy Using an Exogenous Money Supply

Let's consider how a neoclassical economist analyzes the impact of changes in the money supply on the financial markets and the economy. Consider the case of expansionary monetary policy first. If the Fed acts to expand the M1 money supply, then it must choose one of the three following policy actions:

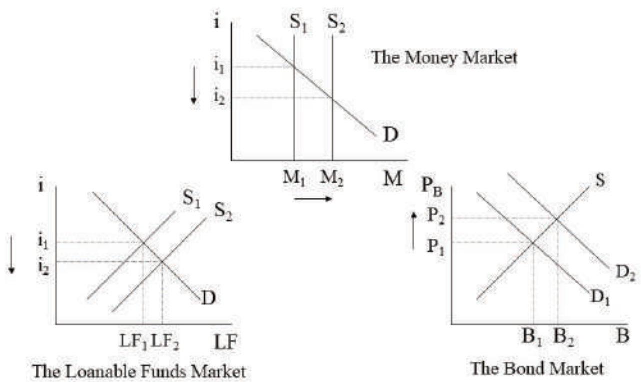
1. Purchase bonds from commercial banks
2. Reduce the discount rate
3. Lower the reserve requirement ratio

Each of these actions will increase the excess reserves available to commercial banks. Using the excess reserves, banks will increase their lending, which will create more checkable deposits and expand the M1 money supply.

The impact of the Fed's increase in the money supply is essentially the same as the impact of an increase in the money supply that commercial banks entirely initiate, as described in Chapter 16. Indeed, the Fed is only able to increase the money supply because it encourages banks to lend more. The Fed's role is different, however, in that it has the legal authority to alter the money supply in a deliberate way and on a massive scale. Each commercial bank, on the other hand, is focused on making a profit and only alters the money supply unintentionally and on a small scale.

The impact on financial markets of the Fed's monetary expansion is shown in Figure 17.8.

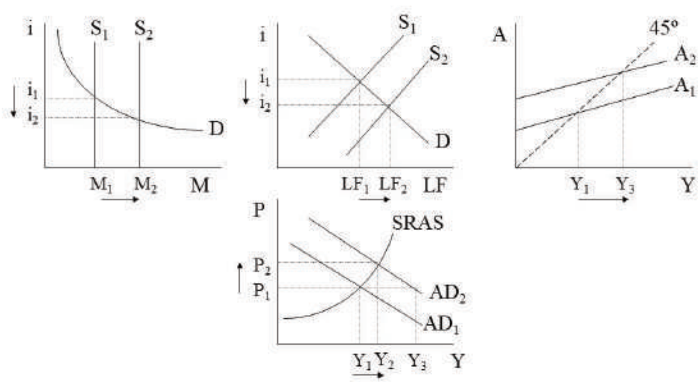
Figure 17.8: The Fed Pursues an Expansionary Monetary Policy



As Figure 17.8 shows, the money supply curve shifts rightward as banks lend more and create more checkable deposits. This shift creates a surplus of money in the money market. That is, firms and households are holding more money than they wish to hold. Consequently, they lend the surplus funds in the loanable funds market, which is equivalent to the purchase of additional bonds in the bond market. In the loanable funds market, the increase in supply creates a surplus of loanable funds. Competition in the loanable funds market drives down the rate of interest to clear that market. Similarly, the higher demand for bonds creates a shortage of bonds in the bond market. Competition in the bond market drives up the price of bonds until the bond market clears. Simultaneous equilibrium thus occurs in all three markets. The Fed's expansionary monetary policy has pushed interest rates down and bond prices up.[5]

The Fed is primarily interested in promoting increased output and employment as well as a stable price level. If it has made the decision to increase the money supply, then it must believe that the economy is operating at less than full employment. Otherwise, a monetary expansion would be purely inflationary. Figure 17.9 shows the impact of an easy money policy in the case of an economy operating below full employment.

Figure 17.9: The Impact of Expansionary Monetary Policy on Investment Spending

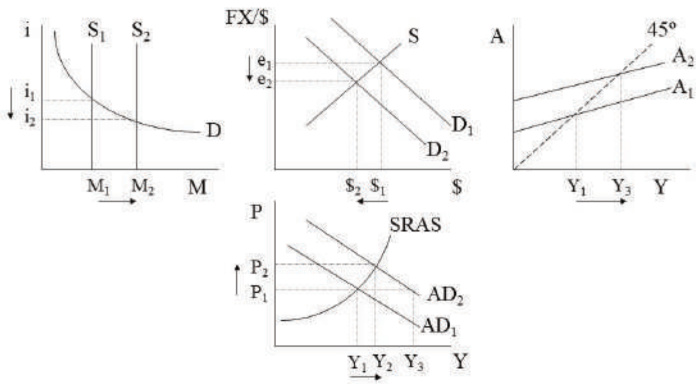


As Figure 17.9 shows, when the supply of money increases, the increased money supply leads to a surplus of money in the money market, which is loaned to borrowers in the loanable funds market. Interest rates fall as described previously. The reduction in interest rates then has an impact on the real economy. The type of aggregate spending that is most sensitive to changes in interest rates is investment spending. When interest rates fall, businesses can obtain loans more cheaply, which they use for the purchase of new capital equipment and new structures (e.g., office buildings, factories, production plants). Home buyers also can obtain mortgage loans more cheaply, which they use to buy new homes. The increase in business investment and residential fixed investment increases aggregate expenditures in the economy causing the A curve to shift upwards in the Keynesian Cross model. The consequence is a rise in the equilibrium level of real GDP.

The level of aggregate demand (AD) also increases. The reader should notice that the rise in AD puts upward pressure on the price level. As the economy approaches full employment, the resulting diminishing returns to labor and the use of less efficient resources pushes unit costs upward, and so firms are forced to raise prices. Therefore, real GDP rises to Y_2, which is not quite as much as the increase to Y_3. The reason is that the rise in the price level prevents aggregate expenditure from rising quite as much due to the wealth effect, the international substitution effect, and the interest-rate effect described in Chapter 13. In any case, the Fed's monetary policy has expanded the level of real GDP, but it has also caused some demand-pull inflation.

The impact on investment spending is not the only impact that the Fed's monetary expansion has on aggregate spending. Figure 17.10 shows a second kind of impact that the Fed's monetary expansion has on the economy.

Figure 17.10: The Impact of Expansionary Monetary Policy on Net Export Spending



In Figure 17.10, we can see that the increase in the money supply puts downward pressure on the rate of interest in the money market as previously described. Foreign investors are likely to be affected when they see that the return on interest-bearing assets in the United States is lower. Specifically, they will be less likely to purchase assets, such as U.S. savings deposits, CDs, and bonds. Because they purchase fewer U.S. interest-bearing assets, they will also purchase fewer U.S. dollars in the foreign exchange market, which are needed to purchase U.S. goods and assets. The reduction in demand for U.S. dollars in the foreign exchange market will cause the foreign exchange value of the U.S. dollar to fall. That is, the U.S. dollar will depreciate.

The depreciation of the U.S. dollar makes U.S. goods relatively cheaper in international commodity markets, which stimulates U.S. exports. The increase in U.S. exports raises aggregate expenditure and boosts real GDP as shown in the Keynesian Cross model. As in the case of a boost to investment spending, the increase in aggregate demand pushes up the price level and prevents real GDP from rising as much as is shown in the Keynesian Cross model. The closer the economy is to full employment, the more the price level will rise and the less the economy will expand in real terms. In general, the Fed's monetary expansion boosts both investment spending and net export spending, allowing the economy to expand but at the cost of producing some demand-pull inflation.

Now let's consider the case of contractionary monetary policy. If the Fed acts to reduce the M1 money supply, then it must choose one of the three following policy actions:

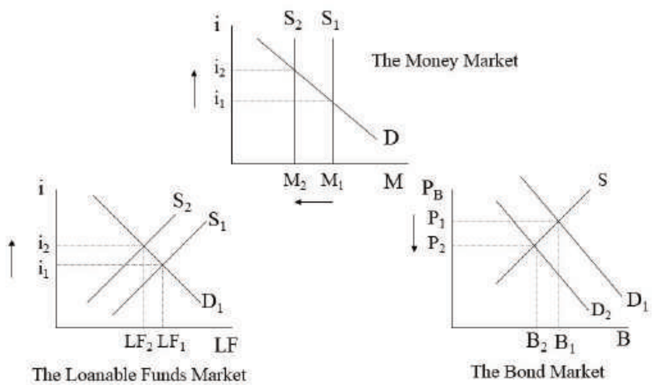
1. Sell bonds to commercial banks
2. Raise the discount rate
3. Raise the reserve requirement ratio

Each of these actions will reduce the excess reserves available to commercial banks. With fewer excess reserves, banks will reduce their lending, which will cause a reduction in the amount of checkable deposits and contract the M1 money supply.

The impact of the Fed's reduction in the money supply is essentially the same as the impact of a decrease in the money supply that commercial banks entirely initiate, as described in Chapter 16. Indeed, the Fed is only able to decrease the money supply because it discourages banks from lending. Again, the Fed's role is different, however, in that it has the legal authority to reduce the money supply in a deliberate way and on a massive scale. Each commercial bank, on the other hand, is focused on making a profit and only decreases the money supply unintentionally and on a small scale.

The impact on financial markets of the Fed's monetary contraction is shown in Figure 17.11.

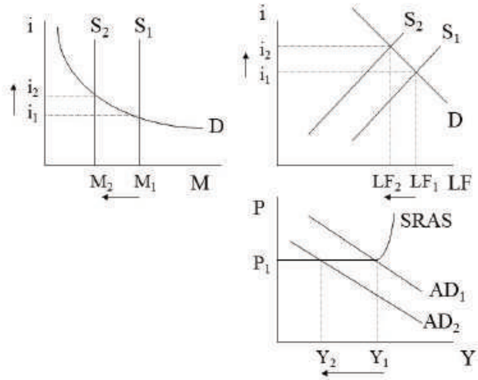
Figure 17.11: The Fed Pursues a Contractionary Monetary Policy



As Figure 17.11 shows, the money supply curve shifts leftward as banks lend less and create fewer checkable deposits. This shift creates a shortage of money in the money market. That is, firms and households are holding less money than they wish to hold. Consequently, they lend fewer funds in the loanable funds market, which is equivalent to the purchase of fewer bonds in the bond market. In the loanable funds market, the reduction in supply creates a shortage of loanable funds. Competition in the loanable funds market drives up the rate of interest to clear that market. Similarly, the lower demand for bonds creates a surplus of bonds in the bond market. Competition in the bond market drives down the price of bonds until the bond market clears. Simultaneous equilibrium thus occurs in all three markets. The Fed's contractionary monetary policy has pushed interest rates up and bond prices down.

The Fed is primarily interested in promoting increased output and employment as well as a stable price level. If it has made the decision to decrease the money supply, then it must believe that the economy is operating at close to full employment and that inflation is a serious concern. Otherwise, a monetary contraction would worsen an already sluggish economy. Figure 17.12 shows the impact of a tight money policy in the case of an economy operating near full employment.

Figure 17.12: The Impact of Contractionary Monetary Policy on Investment Spending



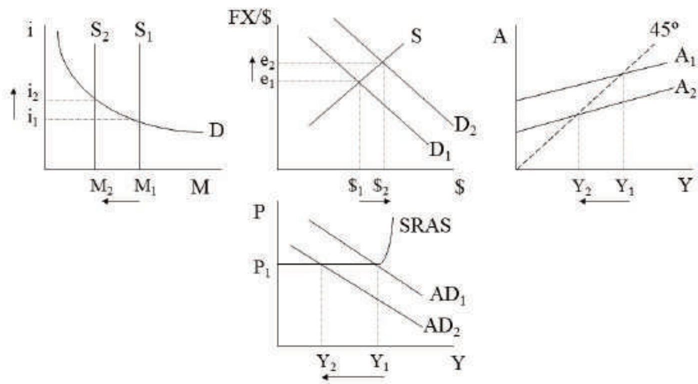
As Figure 17.12 shows, when the supply of money contracts, the reduced money supply leads to a shortage of money in the money market, which leads to less lending to borrowers in the loanable funds market. Interest rates rise as described previously. The rise in interest rates then has an impact on the real economy. Again, the type of aggregate spending that is most sensitive to changes in interest rates is investment spending. When interest rates rise, businesses must pay more to obtain loans for the purchase of new capital equipment and new structures (e.g., office buildings, factories, production plants). Home buyers must also pay more for mortgage loans, which they use to buy new homes. The reduction in business investment and residential fixed investment decreases aggregate expenditures in the economy causing the A curve to shift downwards in the Keynesian Cross model. The consequence is a drop in the equilibrium level of real GDP.

The level of aggregate demand (AD) also decreases. The reader should notice that the drop in AD puts downward pressure on the price level. If prices are sticky in a downward direction, then the price level will remain stable as the level of output and employment fall. Therefore, real GDP falls to Y_2, which is the same reduction in real GDP that occurs in the Keynesian Cross model. The reason that the reduction in real GDP is the same in both the Keynesian Cross and AD/AS models is that the price level is sticky. If the price level does decline, then the wealth effect, the international substitution effect, and the interest-rate effect would cause a movement along the AD curve and partially offset the drop in aggregate spending from the higher interest rates. In any case, the Fed's monetary policy has reduced the level of real GDP.

It might seem surprising that the Fed would ever pursue a contractionary monetary policy. Indeed, it appears that the Fed is trying to engineer a recession! The Fed might pursue a minor contraction, however, if the economy is approaching full employment and a rise in inflation appears to be a likely result. Fed officials might reason that it is better to weaken the economy a little bit than to allow the boom to create so much inflation that a major monetary contraction is required. The former Federal Reserve Chair, Janet Yellen, argued in 2017 that such a **boom-bust policy** should be avoided.

The impact on investment spending is not the only impact that the Fed's monetary contraction has on aggregate spending. Figure 17.13 shows a second kind of impact that the Fed's monetary contraction has on the economy.

Figure 17.13: The Impact of Contractionary Monetary Policy on Net Export Spending



In Figure 17.13, we can see that the decrease in the money supply puts upward pressure on the rate of interest in the money market as previously described. Foreign investors are likely to be affected when they see that the return on interest-bearing assets in the United States is higher. Specifically, they will be more likely to purchase assets, such as U.S. savings deposits, CDs, and bonds. Because they purchase more U.S. interest-bearing assets, they will also purchase more U.S. dollars in the foreign exchange market, which are needed to purchase U.S. goods and assets. The increased demand for U.S. dollars in the foreign exchange market will cause the foreign exchange value of the U.S. dollar to rise. That is, the U.S. dollar will appreciate.

The appreciation of the U.S. dollar makes U.S. goods relatively more expensive in international commodity markets, which reduces U.S. exports. The decrease in U.S. exports lowers aggregate expenditure and reduces real GDP as shown in the Keynesian Cross model. As in the case of a contraction of investment spending, the decrease in aggregate demand puts downward pressure on the price level. If prices are sticky in a downward direction, then the price level will remain stable as the level of output and employment fall. Therefore, real GDP falls to Y_2, which is the same reduction in real GDP that occurs in the Keynesian Cross model. As stated previously, the reason that the reduction in real GDP is the same in both the Keynesian Cross and AD/AS models is that the price level is sticky. If the price level does decline, then the wealth effect, the international substitution effect, and the interest-rate effect would cause a movement along the AD curve and partially offset the drop in aggregate spending from the higher interest rates. In any case, the Fed's monetary policy has reduced the level of real GDP. In general, the Fed's monetary contraction reduces both investment spending and net export spending, allowing the economy to avoid a surge of inflation but at the cost of some reduction in aggregate output and employment.

The Quantity Theory of Money and the AD/AS Model

Neoclassical economists also defend a theory referred to as the Quantity Theory of Money. It is based on an identity known as the Quantity Equation. The **Quantity Equation** identifies a specific relationship between the money supply, the velocity of money, the price level, and the level of real output. The **velocity of money** is the number of times that a unit of the domestic currency (e.g., a U.S. dollar) is spent on average during a given period. For example, if the money velocity is equal to 6, then a dollar bill is spent six times on average during the year. Using this definition of the velocity of money, we can now write the Quantity Equation:

$$MV = PY$$

The Quantity Equation shows that the product of the money supply (M) and the velocity of money (V) is equal to the product of the price level (P) and the level of real output (Y). The product of the price level and the level of real output (PY) is the nominal GDP of the economy. Given the velocity of money, M is the money supply that will support that level of nominal spending. For example, suppose that the money supply is $900 billion and the velocity of money is 6. Then if the price level is 2, the level of real output must be $2,700 billion.

The Quantity Equation is just an identity. It is true given the definition of the variables and does not constitute a theory until we consider how the variables in the equation are causally related. Neoclassical economists have traditionally argued that the velocity of money is relatively stable over time. Hence, the quantity of money is the main determinant of the level of nominal GDP. This view suggests that the Fed has a considerable amount of influence over the economy.

As soon as it is argued that an increase in the money supply causes a rise in nominal GDP given a relatively stable money velocity, the Quantity Equation is transformed into a monetary theory called the **Quantity Theory of Money**. For neoclassical economists, the length of the period under consideration influences the nature of the impact of a money supply change on the economy. For example, in the short run when prices are sticky, a money supply increase (given V) causes an unambiguous rise in the level of real output.

$$M\uparrow \ \overline{V} = \overline{P} \ Y\uparrow$$

The impact of an increase in the money supply in the short run may be depicted as in Figure 17.14.

Figure 17.14: The Impact of Expansionary Monetary Policy in the Short Run

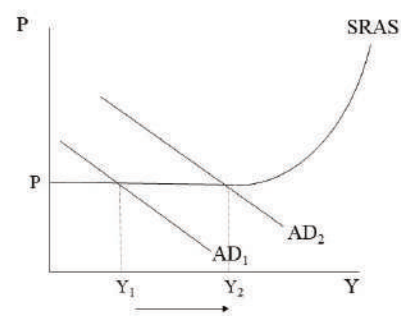


Figure 17.14 shows how the rise in the money supply raises aggregate demand and the level of real output while leaving the price level the same. Similarly, a monetary contraction would lower aggregate demand and the level of real output while leaving the price level the same.

In the long run, the situation is reversed. Neoclassical economists argue that the economy will operate at full employment. That is, the quantities of labor, capital, and land will determine the level of real GDP. Therefore, if the money supply rises in the long run (given V), then the price level will rise and the level of real GDP will remain the same.

$$M\uparrow\ \overline{V} = P\uparrow\ \overline{Y}$$

The impact of an increase in the money supply in the long run may be depicted as in Figure 17.15.

Figure 17.15: The Impact of Expansionary Monetary Policy in the Long Run

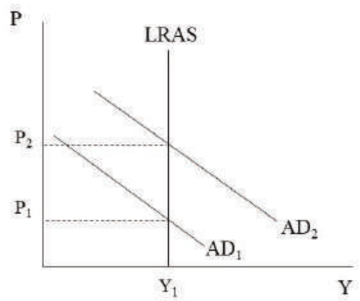


Figure 17.15 shows how the rise in the money supply raises aggregate demand and the price level while leaving the level of real output the same. Similarly, a monetary contraction would lower the price level while leaving the level of real output unchanged.

The Post-Keynesian Approach to Monetary Policy Using an Endogenous Money Supply

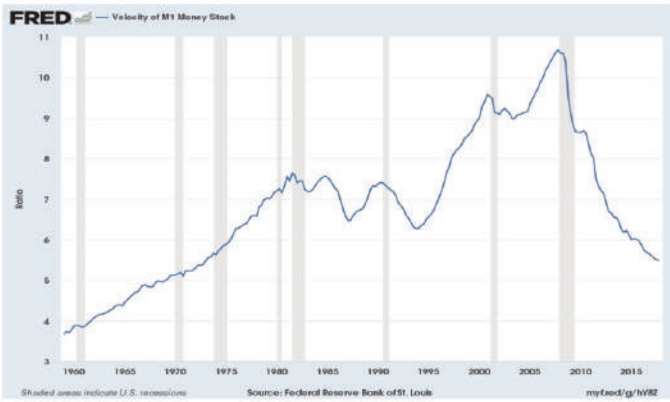
Post-Keynesian economists are critical of the neoclassical approach to monetary policy. Much of the disagreement relates to the assumption that the velocity of money is relatively stable. If the velocity of money is unstable and unpredictable, then changes in the money supply may be offset or intensified when the velocity of money changes. For example, if Fed increases the money supply with the intention of increasing nominal GDP by a given amount, then an unpredictable increase in money velocity may cause an inflationary surge of nominal GDP. On the other hand, if money velocity unpredictably declines, then nominal GDP will not rise as much or may even decline.

Given these questions about the stability and predictability of the velocity of money, it is natural to wonder what the historical evidence suggests. A simple rearrangement of the Quantity Equation allows us to easily calculate the velocity of money.

$$V = \frac{PY}{M}$$

In other words, if we divide nominal GDP by the money supply, we obtain the velocity of money. Figure 17.16 shows the historical pattern of the M1 velocity of money from 1959 to 2017.

Figure 17.16: The Historical Pattern of the M1 Money Velocity, 1959-2017



Source: Federal Reserve Bank of St. Louis, Velocity of M1 Money Stock [M1V], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/M1V, February 11, 2018.

As Figure 17.16 shows, the M1 velocity of money, although increasing steadily, was relatively predictable prior to 1980.[6] This pattern suggests that changes in the money supply could be used to manipulate nominal GDP with a fair amount of accuracy. With sticky prices in the short run, money supply manipulation could be used to influence the level of real GDP, as explained previously. In the long run, however, manipulation of the money supply would alter the price level. In any case, the neoclassical assumption appears to be reasonable prior to 1980.

After 1980, the pattern of money velocity abruptly changes. Money velocity begins to fluctuate in ways that do not appear to follow any predictable pattern. In this environment, altering the money supply with the goal of changing nominal GDP is extremely difficult. Central bankers have no way of knowing how much to change the money supply to alter nominal GDP.

The reasons for the shift are debated. One possibility is that modern information technology has made it easier to transfer funds from non-M1 M2 assets (e.g., savings deposits, MMDAs) to checkable deposits and back again. For example, during the 1990s, we observe a huge jump in the velocity of money. If people began to minimize their M1 money holdings to maximize their interest income from non-M1 M2 assets, then money velocity as defined in this chapter would jump for a given level of nominal GDP. Because it became easier to transfer funds to M1 accounts when they were needed for transactions, many households and firms may have made the decision to hold less M1 money.

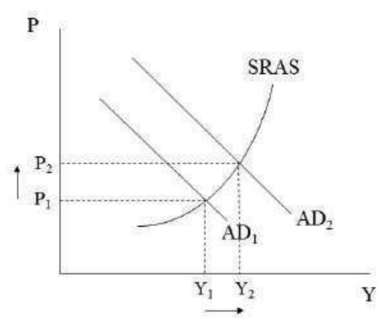
Even though the velocity of money began to change in unpredictable ways after 1980, it is worth noting that during recessions since that time (the shaded areas in Figure 17.16), the velocity of money has always fallen. One possibility is that people wish to hold more money during recessions because they are afraid of job loss. The loss of a job means the loss of income and the inability to pay bills. Holding liquid assets during recessions makes sense then. With an increase in currency holdings and checkable deposits (M), the velocity of money (V) will fall. The reduction in velocity will be even greater due to the contraction of nominal GDP (PY) during the recession. A reference to the formula for money velocity provided above confirms these results.

Due to complications that stem from the use of monetary policy, Post-Keynesians tend to share Keynes's preference for fiscal policy as a means of combatting recessions. The Post-Keynesian objection to neoclassical monetary theory, however, goes beyond the claim that the velocity of money is unstable and unpredictable.

Post-Keynesians also argue that the causal relationship between the money supply and the aggregate economy is exactly the reverse of what neoclassical economists assert.

Before exploring the post-Keynesian alternative monetary theory in detail, let's consider how the neoclassical assertion of an exogenous money supply creates a problem within the AD/AS framework. Suppose that government spending increases in the neoclassical AD/AS model. The result is a rightward shift of the aggregate demand curve as shown in Figure 17.17.

Figure 17.17: The Impact of a Rise in Government Spending in the AD/AS Model (with accommodation from the central bank)



According to neoclassical theory, this rise in AD causes demand-pull inflation as both real output and the price level increase. The increase in both P and Y suggest that nominal GDP (PY) rises. If the velocity of money is relatively stable, which is the neoclassical assumption, then the only way that a higher price level and higher level of real output can occur is if the money supply increases. It would seem then that the only way for this rightward shift of the AD curve to occur is for the commercial banks to accommodate the increased government spending with increased lending. It is conceivable that the banks could accommodate the increased government spending up to a point, without an injection of reserves from the central bank, if the banks possess sufficient excess reserves and are willing to lend. If they lack sufficient excess reserves, then the central bank must accommodate the federal government's desire to spend more.

Suppose that the federal government wishes to increase spending so much that the banks cannot entirely accommodate the desire for an increased money supply by granting more loans. The Fed can directly support the increased government spending through a process referred to as **debt monetization**. When the federal government issues $100 billion in new bonds to support its spending, the central bank can purchase the bonds as an asset and credit the U.S. Treasury's account at the Fed, as shown in Figure 17.18.

Figure 17.18: The Fed Monetizing the Federal Debt

Figure 17.18 (a): The Fed buys government bonds from the U.S. Treasury

Assets	Liabilities + Bank Capital
U.S. Treasury Securities +$100 billion	U.S. Treasury deposits +$100 billion

Figure 17.18 (b): The federal government issues bonds and sells them to the Fed

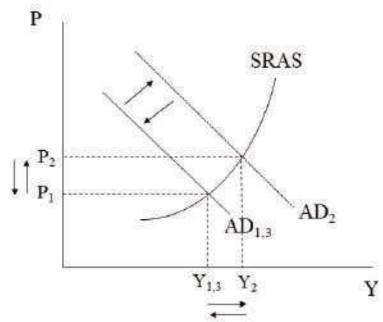
Assets	Liabilities + Bank Capital
U.S. Treasury deposits +$100 billion	U.S. Treasury Securities +$100 billion

As Figure 17.18 shows, the federal government's assets increase when its Treasury deposits at the Fed increase, but its liabilities also increase because it has issued new debt. This increase in Treasury deposits directly adds to the money supply, and the Fed has monetized the debt. In this way, the money supply increase is endogenously related to government spending. That is, rather than the money supply representing an **exogenous variable** that changes for reasons not explained in our model, the money supply represents an **endogenous variable** that changes for reasons explained within our model.

This example shows that the neoclassical commitment to an exogenous money supply makes it difficult to uphold the claim that higher government spending will shift the aggregate demand curve to the right. It is still possible for the government to increase its spending with a constant money supply and money velocity, but only a couple scenarios can explain how such a change occurs.

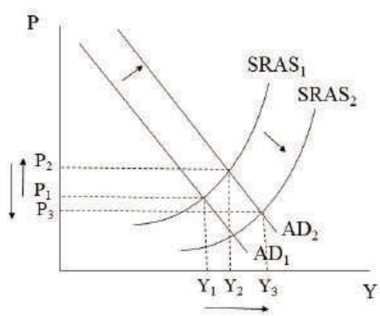
If the government, for example, increases spending, it will increase the production of certain commodities and their prices. With the money supply and money velocity constant, however, a reallocation of spending must occur. That is, spending on other commodities for consumption and investment will decline, causing their production levels and prices to fall. The AD curve does not change since these factors offset one another in the calculation of nominal GDP, and overall PY does not change as shown in Figure 17.19.

Figure 17.19: The Impact of a Rise in Government Spending in the AD/AS Model (with no accommodation from the central bank and consumer spending and investment spending falling)



Another possibility is that the Fed does not accommodate the government spending increase with a money supply increase, but the reallocation of spending occurs differently. For instance, the increased government spending that pushes up the prices and production of some commodities comes at the expense of spending in input markets. The reduction in spending in input markets could cause the prices of inputs to fall so much that unit costs for firms decline. A reduction in unit production costs shifts the short run aggregate supply (SRAS) curve to the right as explained in Chapter 13. This rightward shift of SRAS puts downward pressure on the price level even though it leads to a higher level of real output. If the fall in the price level is sufficiently large, then it might offset the rise in the prices of government-purchased commodities and the rise in real output. Overall, the economy will expand even though the money supply and velocity of money are constant. This possibility is represented in Figure 17.20.

Figure 17.20: The Impact of a Rise in Government Spending in the AD/AS Model (with no accommodation from the central bank and a decline in spending in input markets)



The outcome represented in Figure 17.20 is very unlikely to occur. It requires the reallocation of spending to affect the input markets the most. It also requires a very steep AD curve so that the shift of the SRAS curve has a relatively much larger impact on the price level than on the level of real output. In this way, real output can expand even with nominal GDP remaining constant. This case represents a major economic boom without the Fed initiating or even accommodating it. Because this outcome is unlikely to occur, it still makes sense to conclude that a rightward shift of the AD curve is very unlikely to occur following an increase in government spending without some accommodation from the central bank.

It might seem that inflation must require accommodation from the commercial banks and the central bank. That conclusion only follows in the case of demand-pull inflation. Cost-push inflation can occur without a money supply increase. This case is represented in Figure 17.21.

Figure 17.21: The Impact of Rising Input Costs in the AD/AS Model (with no accommodation from the central bank)

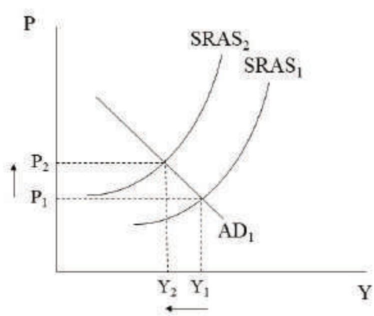


Figure 17.21 shows how a rise in input costs (e.g., wages, oil prices) leads to a leftward shift of the SRAS curve. The result is a rise in the general price level and a reduction in the level of real output. In this case, the fall in real output

implies that the rising price level can occur even if the money supply and velocity of money remain constant.

$$\overline{MV} = P\uparrow \quad Y\downarrow$$

This analysis shows that cost-push inflation does not require accommodation from the banking system and the central bank.

We have seen how the assumption of an **exogenous money supply** creates problems for the traditional neoclassical analysis of the macroeconomy using the AD/AS model. Let's take a closer look at why Post-Keynesian economists argue that the supply of money should be treated as an endogenous variable in macroeconomic theory. Post-Keynesian economists typically argue that owners of different factors of production are involved in a conflict over the distribution of income. Workers, for example, demand higher wages and form labor unions to apply pressure on employers. Firms, on the other hand, use their market power to maintain markups of product prices over unit labor costs. A constant struggle ensues that can generate cost-push inflation and a **wage-price inflationary spiral**. This relationship can be captured best using the following equation:[7]

$$P = m \cdot \frac{wL}{Q}$$

In the equation above, the price of a product (P) is equal to a markup (m) over the per unit labor cost of production. In this case, the markup is multiplied by the ratio of total labor cost (wL) to total output. Total labor cost is the wage rate (w) times the number of units of labor hired (L). When this total wage bill is spread across the number of units produced (Q), we obtain the labor cost per unit (wL/Q).

When workers demand and win wage increases, firms raise their product prices since they desire a constant markup over unit labor costs. When product prices increase, workers recognize that their real wages (w/P) are eroded, which brings their purchasing power back to the original level. Workers thus demand higher nominal wage increases, which continues the process. One should not assume that workers are the guilty party initiating the wage-price inflationary spiral. Firms might initiate the cycle with an initial increase in the markup over unit labor costs, which raises prices and per unit profits. Workers may then respond with the demand for higher money wages.

The upward pressure on prices means that firms and consumers will have a more difficult time making purchases and will need to obtain more loans to make purchases. They will also need to hold more checkable deposits and currency to engage in desired transactions. Post-Keynesian economists argue that the commercial banks and the central bank will accommodate the desire for additional loans and money balances. That is, the assumption that the money supply is determined endogenously in response to these events carries very different consequences for the financial markets than what we observed when an exogenous money supply was assumed. Figure 17.22 shows the impact on the financial markets of an increased demand for money and loans in the context of an **endogenously determined money supply**.

Figure 17.22: The Impact of the Wage-Price Inflationary Spiral on the Financial Markets (according to Post-Keynesian theory)

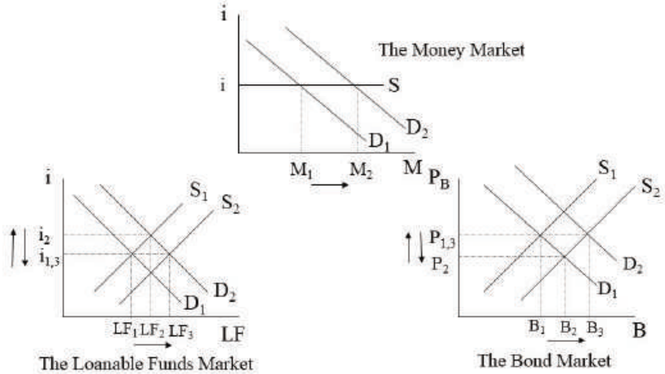


Figure 17.22 shows that the rate of interest does not change when the central bank completely accommodates the higher demand for money. As the demand for money rises, the slightest rise in the rate of interest leads to increased commercial bank lending and increased open market purchases by the Fed, which provides banks with more reserves to encourage bank lending. The money supply curve is perfectly elastic as a result. At the same time, the higher demand for loans in the loanable funds market is completely accommodated as the Fed provides banks with more reserves, and the banks lend more. The rise in the supply of loanable funds keeps the rate of interest from rising in the loanable funds market. As the demand for loans rises, the counterpart of this change in the bond market is a higher supply of bonds, which puts downward pressure on bond prices. Because banks accommodate securities dealers' higher demand for

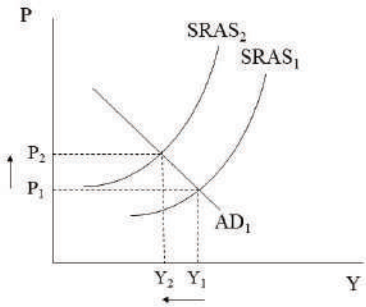
checkable deposits by purchasing their securities, the demand for bonds rises too. This accommodating stance keeps bond prices from falling. Hence, simultaneous equilibrium in all three markets occurs, and interest rates and bond prices remain unchanged.

The constant rate of interest suggests that the cost of borrowing does not change. Hence, aggregate investment spending and consumer spending, which are sensitive to interest rate changes will not change. We should not expect an aggregate demand shift to occur. Even though firms and households are borrowing more, they are doing so in response to a higher price level. It does not represent an increase in the demand for real goods and services. Another way to think about why the AD curve does not shift is to consider the impact on the velocity of money. The higher demand for money has a lower money velocity as its counterpart. Even though the money supply rises, the velocity of money falls, and so AD does not change overall. Nevertheless, the rise in per unit costs does cause a leftward shift of the SRAS curve, and so stagflation results.

$$M\uparrow \quad V\downarrow = P\uparrow \quad Y\downarrow$$

Figure 17.23 shows a Post-Keynesian result that involves a wage-price inflationary spiral and recession with accommodation from the Fed and the commercial banks.

Figure 17.23: The Impact of the Wage-Price Inflationary Spiral (with accommodation from the central bank)



In Figure 17.23 the Fed and the banks are accommodating the higher demand for money and loans resulting from the wage-price inflationary spiral. A lending boom takes place within an inflationary economy where output and employment are falling. The lending allows firms and households to continue spending and driving up the price level even as the economy contracts.

The contrast between this Post-Keynesian analysis and the neoclassical analysis should be clear. If the Fed controls the money supply and does not respond to demands of borrowers and banks, then an exogenous money supply increase will push interest rates down and stimulate investment spending and consumer spending. The result will be a rise in aggregate demand and an economic boom capable of producing higher levels of real output and a higher price level as shown in Figure 17.9. The Post-Keynesian analysis, on the other hand, reveals why the monetary expansion of the 1970s was incapable of pulling the economy out of recession even as it produced high rates of inflation.

A Marxian Theory of Fiat Money and its Relationship to U.S. Economic History

In Chapter 4, the Marxian theory of money was developed. In that theory, money serves as the universal expression of socially necessary abstract labor time (SNALT). Each commodity in circulation is equated with the specific quantity of the money commodity that contains the same amount of SNALT. As paper symbols began to circulate as a representative of the money commodity, each commodity acquired a paper money price. Because the paper symbols were convertible into the money commodity (e.g., gold), the paper value of a commodity represented the exact quantity of the money commodity that contained an equivalent amount of SNALT as the commodity. Convertible paper money, therefore, does not present any special challenges for **Marxian monetary theory**.

As suggested in Chapter 4, the greater challenge to Marxian monetary theory is the existence of inconvertible paper money or fiat money. Because fiat money is not directly convertible into a commodity, it appears that the paper itself possesses value. The problem is that the SNALT required to produce the paper symbols is negligible. Therefore, it does not make sense to compare the labor time required to produce a house with the labor time required to produce $200,000 worth of fiat paper. To make such a comparison would reveal that the labor time required to produce the house far exceeds the labor time required to produce the paper money. The discrepancy seems to suggest that a **Marxian theory of fiat money** is doomed.

How can we determine the specific quantity of fiat paper for which a commodity will exchange if commodities exchange at their values? Within a system of simple commodity circulation, commodities are constantly being exchanged for money and money is constantly being exchanged for commodities. Commodity circuits with the form C-M-C' are used to represent these patterns of sale and purchase. The question we are asking is the following: How much money (M) is needed to complete the circuit when the exchange of equivalent values (C for C') is assumed? In the case of commodity money or convertible paper money, the answer is simple. We just use the amount of commodity money or convertible paper money that represents the same amount of SNALT as the commodities being sold and purchased. That is, a unit of commodity money requires so much SNALT for its production. We simply adjust the quantity of the money commodity so that its value is equivalent to the values of the commodities being exchanged. With fiat paper money, we cannot rely on that solution because, as already argued, the SNALT required for its production is negligible and so it cannot provide the solution if one exists.

The way to resolve the problem is to recognize that fiat paper money continues to represent a claim to other commodities in circulation. Although it cannot be redeemed at a bank for a specific quantity of gold, it can be easily exchanged for commodities. The answer then requires that we reflect on the entire world of circulating commodities. Consider a period during which the circulating paper money supply (M) is given.[8] On average each unit of paper money is spent a specific number of times, which we have called the velocity of money (V). The fiat money price of each commodity represents some fraction of this total expenditure represented as MV. What is the fraction of MV that a specific commodity will command as a price when commodities exchange at their values? It is determined by the SNALT required for its production. Given this reasoning, the fiat money value of a commodity j will equal the following:

$$P_j = \frac{L_j}{\Sigma_{i=1}^{n} L_i} MV$$

In this equation, L_j represents the SNALT required to produce commodity j and ΣLi represents the aggregate SNALT embodied in all n circulating commodities.[9] P_j is the fiat money price or paper value of the commodity j. Only when commodity prices are determined according to this formula will all commodities exchange according to the quantities of SNALT required for their production.

The formula for the fiat money price of commodity j reveals that the paper value of a commodity depends on several factors:

1. It depends positively on the SNALT required for its production (L_j). If the SNALT required to produce the commodity rises, then the paper value of the commodity will increase, other factors held constant (and vice versa).
2. It depends negatively on the aggregate SNALT embodied in all circulating commodities. Other factors held constant, if the economy expands and more SNALT is required to produce all

circulating commodities, then the paper value of the commodity will fall. That is, the commodity represents a smaller fraction of the aggregate labor time embodied in commodities and so it commands a smaller part of the **effective money supply** (MV). The opposite occurs if the economy contracts.

3. It depends positively on the supply of paper money (M). If the money supply rises, other factors held constant, then the fiat paper price of commodity j will increase (and vice versa). In this case, the fraction of the aggregate SNALT devoted to this commodity is applied to a larger effective money supply (MV).
4. It depends positively on the velocity of money (V). If the velocity of money increases, then the fiat money price of commodity j will increase (and vice versa). In this case as well, the fraction of the aggregate SNALT devoted to this commodity is applied to a larger effective money supply (MV).

It turns out that this expression for the fiat money price of commodity j is consistent with the Quantity Equation that neoclassical economists describe. To understand the relationship between the fiat money price equation and the Quantity Equation, try adding up all the fiat money prices in the entire economy. Because n commodities are in circulation, we can write the following equation:

$$P_1+P_2+P_3+...+P_n=\frac{L_1}{\Sigma_{i=1}^n L_i}MV+\frac{L_2}{\Sigma_{i=1}^n L_i}MV+\frac{L_3}{\Sigma_{i=1}^n L_i}MV+...+\frac{L_n}{\Sigma_{i=1}^n L_i}MV$$

$$P_1+P_2+P_3+...+P_n=(\frac{L_1}{\Sigma_{i=1}^n L_i}+\frac{L_2}{\Sigma_{i=1}^n L_i}+\frac{L_3}{\Sigma_{i=1}^n L_i}+...+\frac{L_n}{\Sigma_{i=1}^n L_i})MV$$

$$P_1+P_2+P_3+...+P_n=(\frac{L_1+L_2+L_3+...+L_n}{\Sigma_{i=1}^n L_i})MV$$

$$P_1+P_2+P_3+...+P_n=(\frac{\Sigma_{i=1}^n L_i}{\Sigma_{i=1}^n L_i})MV$$

$$P_1 + P_2 + P_3 + ... + P_n = MV$$

This derivation shows that the sum of all prices of circulating commodities is equal to the effective fiat paper money supply. Because some of the prices listed individually in this equation are for the same commodity, we could write the equation using the products of prices and quantities. If we only include prices that are uniquely associated with one commodity and multiply by the quantity in circulation, then we obtain the equation below:

$$P_1'Y_1 + P_2'Y_2 + P_3'Y_3 + ... + P_k'Y_k = MV$$

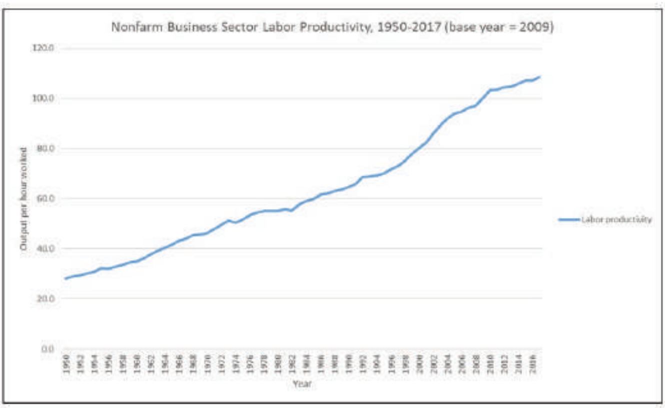
In this equation, P_i' refers to the price of commodity i where all the other prices of commodity i have been eliminated to allow for multiplication by the quantity of that commodity in circulation (Y_i). In other words, the equation reduces to the Quantity Equation:

$$PY = MV$$

Like in neoclassical theory, the Quantity Equation is a simple identity. In Marxian economics, this identity becomes a theory when the prices of the commodities are explained using their individual labor values, the aggregate labor time required for their production, the supply of paper money, and the velocity of money.

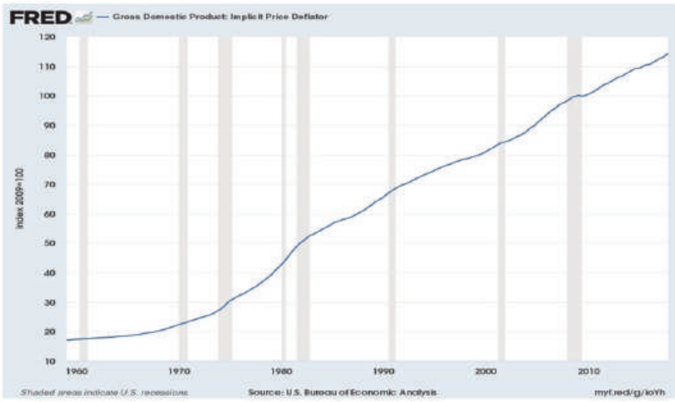
The theory provides us with a way of interpreting U.S. macroeconomic history. For example, it is widely recognized that capitalist development has led to an increase in productivity in many sectors of the economy. Increases in productivity allow production to increase more rapidly than production cost. That is, total output per dollar of production cost rises. The flipside of a rise in productivity then is a reduction in production cost per unit of output. In a competitive economy, a reduction in per unit cost should also causes prices to fall as firms compete and profits are pushed down to a level that is consistent with the general rate of profit in the economy. Figures 17.24 and 17.25 show how average labor productivity in the U.S. and the general price level have changed since the 1950s.

Figure 17.24: Nonfarm Business Sector Labor Productivity, 1950-2017, all employed persons



Source: U.S. Bureau of Labor Statistics, Division of Major Sector Productivity. Data reflect press release of February 1, 2018. Full report: www.bls.gov/news.release/prod2.nr0.htm

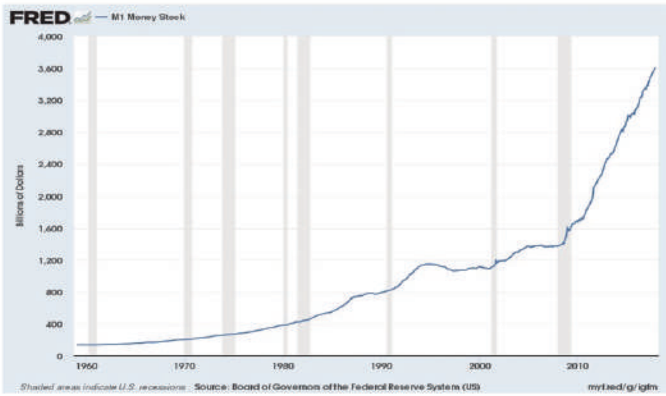
Figure 17.25: The GDP Deflator, 1959-2017, seasonally adjusted, Index 2009 = 100



Source: U.S. Bureau of Economic Analysis, Gross Domestic Product: Implicit Price Deflator [GDPDEF], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/GDPDEF, February 13, 2018.

As Figure 17.24 shows, the average productivity in the U.S. has risen considerably since 1950. The general price level has also risen as shown in Figure 17.25 for the period 1959-2017. This finding seems to contradict the argument that rising productivity lowers per unit cost and prices in a competitive market economy. How do we resolve this paradox? To resolve the paradox, we need only consider the formula for the price of commodity j. When productivity increases, the SNALT required to produce commodity j falls. Other factors the same, the fiat money price of commodity j will fall. Other factors are not the same, however, because the money supply has grown enormously since 1959 as shown in Figure 17.26.

Figure 17.26: The M1 Money Supply, 1959-2017, seasonally adjusted, in billions of dollars



Source: Board of Governors of the Federal Reserve System (US), M1 Money Stock [M1SL], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/M1SL, February 13, 2018.

As the money supply rises, the price of commodity j increases, which shows the inflationary impact of a rise in the money supply. Therefore, the upward pressure on the price due to the increased money supply has offset the downward pressure on the price due to increased productivity. If productivity had not changed at all during this period, then the price of commodity j would have increased even more.

The Marxian theory of fiat money also allows us to respond to a common criticism of Marxian economics. It has been argued that the working class has gained tremendously within capitalist societies as the standard of living has risen. That is, the **real wage** or the commodity bundle that workers generally consume has increased over time. Because this increase in the real wage has occurred, critics argue, workers are better off, and the suggestion that capitalism would lead to an "increasing misery of the proletariat" is unfounded.[10]

Richard Wolff and Stephen Resnick, however, demonstrate that a rising real wage may occur even as the rate of exploitation rises.[11] To understand their point, suppose that NL represents the **necessary labor** time expended in the production of commodity j. The reader should recall that necessary labor time refers to the time required to produce a value equivalent to the wage that the worker is paid. The worker must be paid enough for her labor-power to purchase the means of subsistence for the worker and her dependents. If q represents the real wage (expressed in terms of physical units of commodities that workers require for the reproduction of their labor-power), and e represents the **SNALT required per unit of commodity** in the commodity bundle, then we can write the following equation:

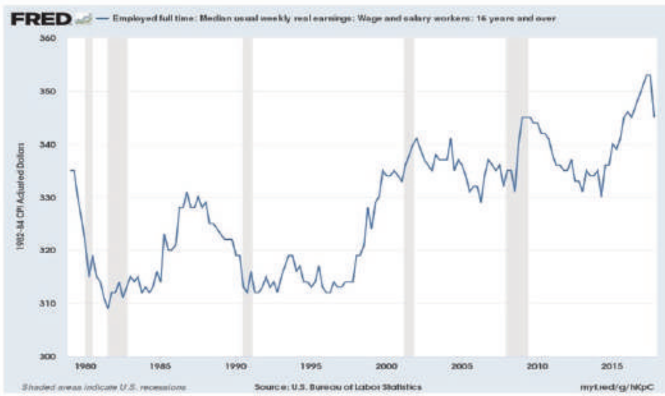
$NL = eq$

Because the worker requires many different commodities, and each has its own labor requirement, eq represents the product of two vectors or:

$eq = e_1q_1 + e_2q_2 + e_3q_3 + ... + e_mq_m$

In this case, the worker consumes m commodities as part of the required commodity bundle. We can now see how American capitalist development is consistent with a rising rate of exploitation even as the real wage has risen. As productivity has risen, the SNALT required for each commodity in the worker's commodity bundle has fallen. At the same time, workers have been able to benefit from this increased productivity as the socially acceptable real wage has increased since 1979 as shown in Figure 17.27.

Figure 17.27: Median Weekly Real Earnings for Full-Time Wage and Salary Workers, 1979-2017, seasonally adjusted, 1982-1984 dollars

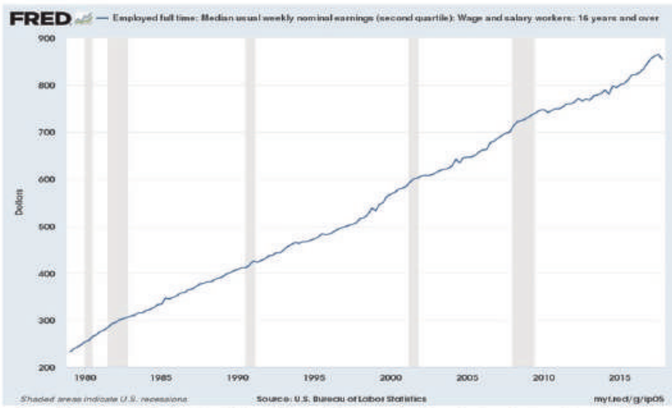


Source: U.S. Bureau of Labor Statistics, Employed full time: Median usual weekly real earnings: Wage and salary workers: 16 years and over [LES1252881600Q], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/LES1252881600Q, February 13, 2018.

Therefore, each q has increased even as each e has fallen. If the productivity increases (and the corresponding reductions in e) are relatively larger than the increases in the real wage, the necessary labor must fall. Given the length of the working day, the surplus labor thus increases, and the rate of exploitation (S/V or SL/NL) rises, as Wolff and Resnick suggest.

The reader might object, however, to this analysis because a reduction in necessary labor time should mean that the **money wage** (or **nominal wage**) has fallen. After all, if it costs less to purchase the commodity bundle due to falling per unit labor values even though the quantities purchased have risen, then less money should be required. U.S. economic history suggests the exact opposite about the overall movement of money wages, which have increased considerably since 1979 as shown in Figure 17.28.

Figure 17.28: Median Weekly Nominal Earnings for Full-Time Wage and Salary Workers, 1979-2017, seasonally adjusted, in dollars



Source: U.S. Bureau of Labor Statistics, Employed full time: Median usual weekly nominal earnings (second quartile): Wage and salary workers: 16 years and over [LES1252881500Q], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/LES1252881500Q, February 13, 2018.

With rising real wages and rising money wages, it seems that Marxian economists cannot defend the claim that workers have been made worse off within capitalist societies. They are not worse off in an absolute sense because of the rise in real wages. They do not appear worse off in a relative sense because of the rise in money wages. The latter suggestion contains a flaw, however, because it ignores the inflationary impact of a rise in the money supply. Looking again at the formula for the price of commodity j, it should be clear that it is possible to multiply any labor time magnitude by the ratio of the effective money supply to the aggregate SNALT embodied in all circulating commodities to obtain a fiat money price. This ratio is a magnitude that Marxian economists refer to as the **monetary expression of labor time (MELT)**:

$$MELT = \frac{MV}{\Sigma_{i=1}^{n} L_i}$$

Hence, the price of commodity j can be written in the following way:[12]
$P_j = MELT \cdot L_j$

We can use the MELT to calculate the **variable capital (v)** as follows:
$v = MELT \cdot NL$

Now consider all the relevant factors that have changed since 1950 in the United States. Productivity has increased, which has caused the real wage to rise and the labor values of commodities to fall. The net change has been a reduction in the necessary labor time because the productivity increases had a relatively larger impact on the labor values than on the real wage. At the same time, the money supply has increased enormously, which has caused the MELT to rise substantially. The variable capital has thus increased, but the increase represents an inflationary increase rather than a redistribution of the new value produced during the workday between capitalists and workers. In other words, the **surplus labor (SL)** time has increased, reflecting a rise in the rate of exploitation, but the surplus value (S) has also become inflated due to the increase in the MELT.
$S = MELT \cdot SL$

It should be clear that the surplus value has risen more than the variable capital. Inflation has caused a rise in both measures due to inflation, but the redistribution from the fall in necessary labor time and the rise in surplus labor time has caused the surplus value to rise relatively more than the variable capital has risen.

Before we conclude, it is also helpful to think about the meaning of the reciprocal of the MELT. Its reciprocal is what we might call the **value of money (VOM)**. It shows us how much SNALT is represented with one unit of currency.

$$VOM = \frac{\Sigma_{i=1}^{n} L_i}{MV}$$

Hence, when the MELT rises, the value of money falls. During periods of price inflation, a reduction in the value of money occurs, and so this result is consistent with our intuition.

Monetary Policy Tools within the Context of the Marxian Theory of Financial Markets

In this section, we will consider how a Marxian economist might analyze the Fed's use of its three primary monetary policy tools. To do so, we will return to the analysis of financial markets that was introduced in Chapter 16. Recall the definition of the general rate of interest (i_g) that was provided in that chapter.

$$i_g = \frac{\Delta M}{M} = \frac{B_O + \Delta B}{B_L + (1-R)D} = \frac{\frac{B_O}{B} + p}{\frac{B_L}{B} + \frac{(1-R)D}{B}}$$

It should be immediately apparent that one of the monetary policy tools has already been included in the definition. The reserve requirement ratio (R) was used to determine the amount of excess reserves that a bank can lend. The bank's total excess reserves are B_L+(1-R)D.

Clearly, if the central bank reduces R, then the bank will have more excess reserves that it can lend. This increase in excess reserves will reduce the general rate of interest because a lower interest rate will be sufficient to generate the general rate of profit on bank capital. The reader should recall that the market rate of interest will tend to fall towards the level of the general rate of interest. As soon as the general rate of interest falls below the market rate of interest, capital will flow out of industry and into finance. The increased lending will push the market rate of interest down until it equals the general rate of interest. The conclusion that the market rate of interest will fall when the reserve requirement ratio is reduced is perfectly consistent with the result that neoclassical economists reach. In neoclassical economics, the reduction in the reserve requirement leads to a shift of the money supply curve to the right, which pushes down the market rate of interest in the money market. In Marxian economics, the reduction in the reserve requirement ratio raises the profitability of banking, which leads to capital inflows and pushes down the market rate of interest as lending increases. The results are similar, but these economists reach their conclusions using different theories.

The opposite case involves the central bank increasing the reserve requirement ratio. In this case, the bank will have fewer excess reserves for lending. The general rate of interest will then be higher because a higher interest rate will be needed to ensure the general rate of profit on bank capital. As the general rate of interest rises above the market rate of interest, capital will flow out of finance and into industry. The consequent reduction in bank lending will lead to a higher market rate of interest. The market rate of interest will continue to rise until it equals the general rate of interest. This result is also consistent with the neoclassical conclusion that an increase in the reserve requirement will raise the market rate of interest.

Let's next consider how open market operations may be incorporated into the Marxian analysis of financial markets. That is, if the central bank buys securities from banks or sells securities to banks, how will that alter the analysis that we have explored thus far? If the central bank is going to be involved in these transactions with commercial banks, then it makes sense to consider how banks acquire securities in the first place.

Consider Table 16.3 that illustrated how the general rate of interest ensures an equal rate of profit in both the industrial sector and the financial sector. In that example, the banks made loans to the industrial capitalists using their excess reserves of $195,000. Industrial capitalists used the borrowed capital of $195,000 to purchase means of production and labor-power. The commodities produced were sold for a profit, and a share of it was given to the bankers in the form of interest.

Now consider how the situation would be different if the bankers sell government bonds to the central bank at the start of this same period. That is, suppose that the central bank decides to purchase government securities in the amount of $97,500 from the commercial banks in an open market purchase. The impact on the central bank's balance sheet is shown in Figure 17.29.

Figure 17.29: The Impact of an Open Market Purchase of Securities on the Central Bank's Balance Sheet

Assets		Liabilities + Bank Capital	
Securities	+\$97,500	Reserves	+\$97,500

Figure 17.29 shows that the central bank acquires an asset in the form of government securities and pays the banks for the bonds with an increase in their reserves. Once the central bank takes the securities off the banks' balance sheets, the banks once again have those excess reserves to lend. In other words, the central bank's purchase of \$97,500 of securities gives the banks new excess reserves, which they can then lend to the industrial capitalists. Table 17.1 shows how the borrowed capital in the industrial sector expands by 1.5 times due to the open market purchase and the additional bank lending that stems from it.

Table 17.1: An Example with Central Bank Purchases of Securities (government bonds)

Industrial Sector		Financial Sector	
Industrial Rate of Profit	25.00%	Financial Rate of Profit	25.00%
Non-Borrowed Capital ($)	305,000	Bank Operating Capital ($)	5,000
Borrowed Capital (Loan Capital) ($)	292,500	Bank Loan Capital ($)	35,000
Total Productive Capital ($)	597,500	Total Deposits ($)	200,000
Total Profit ($)	149,375	Reserve Requirement Ratio	20.00%
Interest Paid to Banks ($)	15,000	**Total Bank Capital ($)**	40,000
Total Annual Social Product ($)	746,875	**General Rate of Interest (Prime Rate)**	≈5.13%
		Loan Interest Received or Gross Bank Profit ($)	15,000
Aggregate Profits ($)	159,375	**Net Bank Profit ($)**	10,000
Total Social Capital ($)	637,500	Open Market Fed Purchases ($)	97,500
General Rate of Profit (Gross)	25.00%		
Interest Paid to Central Bank ($) (interest from government bonds)	5,000		

The increase in borrowing in the industrial sector causes the total productive capital in that sector to expand. Consequently, the total profit and the total social product increase. This change reflects the expansionary impact of open market purchases on the economy. At the same time, a lower rate of interest is necessary to ensure the general rate of profit on bank capital since the bank is lending more than previously. In other words, the general rate of interest has fallen:

$$\dot{i}_g = \frac{\Delta M}{M} = \frac{B_O+\Delta B}{F_P+B_L+(1-R)D} = \frac{\frac{B_O}{B}+p}{\frac{F_P}{B}+\frac{B_L}{B}+\frac{(1-R)D}{B}}$$

In this new formula for the general rate of interest, F_P represents **Federal Reserve open market purchases of securities**. These purchases allow the banking system's excess reserves to increase, which increases lending. Therefore, the rate of interest that is necessary to ensure the general rate of profit on bank capital is lower. Table 17.1 shows that the general rate of interest has fallen to

5.13% (compared with 7.69% prior to the open market purchases). This reduced rate of interest allows the banking sector to appropriate the same amount of interest as before, which ensures a 25% rate of profit. Because the government must also pay interest to the central bank, the interest owed may be calculated as the product of the general rate of interest and the face value of the securities to obtain $5,000 (= 5.13% times $97,500). Industrial capitalists now owe $15,000 in interest payments to the commercial banks due to their expanded borrowing, and the government owes $5,000 in interest payments to the central bank.

This example shows how open market purchases push down the general rate of interest and the market rate of interest while promoting an economic expansion. The process may be reversed if the central bank sells securities it owns to the commercial banks. In that case, F_P will decline and the general rate of interest and the market rate of interest will rise. In that case, the banks will lose reserves and will contract their lending.

Finally, let's consider how central bank discount lending affects the economy and the general rate of interest. In this example, we will assume that the open market purchase described previously has occurred. If the central bank next extends discount loans to commercial banks, then the banks acquire new excess reserves. Assume that the central bank lends $97,500 in discount loans to banks. The banks then grant this entire amount as additional loans to industrial capitalists bringing the total borrowed capital to $390,000. The change to the central bank's balance sheet, which includes the open market purchase described previously, is shown in Figure 17.30.

Figure 17.30: The Impact of an Open Market Purchase and Discount Lending on the Central Bank's Balance Sheet

Assets		Liabilities + Bank Capital	
Securities	+$97,500	Reserves	+$195,000
Discount Loans	+$97,500		

Figure 17.30 shows that the central bank has acquired discount loans as a new asset in addition to the securities previously acquired through the open market purchase of bonds from the banks. The discount loans are granted in the form of reserves, which increases the reserve accounts of banks by $195,000. It is assumed that this entire amount is loaned to the industrial capitalists. The additional $97,500 of discount loans has led to an expansion of the industrial sector by this amount due to increased commercial bank lending stemming from the discount loans. Table 17.2 shows how the borrowed capital expands in the industrial sector when bank lending rises following the granting of the discount loans.

Table 17.2: An Example with Central Bank Discount Lending (A Low Discount Rate)

Industrial Sector		Financial Sector	
Industrial Rate of Profit	25.00%	Financial Rate of Profit	25.00%
Non-Borrowed Capital ($)	305,000	Bank Operating Capital ($)	5,000
Borrowed Capital (Loan Capital) ($)	390,000	Bank Loan Capital ($)	35,000
Total Productive Capital ($)	695,000	Total Deposits ($)	200,000
Total Profit ($)	173,750	Reserve Requirement Ratio	20.00%
Interest Paid to Banks ($)	16,950	**Total Bank Capital ($)**	40,000
Total Annual Social Product ($)	868,750	**General Rate of Interest (Prime Rate)**	≈4.35%
		Loan Interest Received or Gross Bank Profit ($)	16,950
Aggregate Profits ($)	183,750	**Net Bank Profit ($)**	10,000
Total Social Capital ($)	735,000	**Interest Paid to the Central Bank ($) (interest from discount loans)**	1,950
General Rate of Profit (Gross)	25.00%	Open Market Fed Purchases ($)	97,500
Interest Paid to Central Bank ($) (interest from government bonds)	4,237.50	Discount Loans	97,500
		Discount Rate	2.00%

With the expansion of the borrowed capital in the industrial sector, the productive capital and the profit also increase. The total social product also rises. The discount lending has led to an expansion of the economy, which suggests that it is an expansionary monetary policy.

It is also worth considering the impact that the discount lending has on the rate of interest. The general rate changes for two reasons due to the granting of the discount loans as shown below:

$$i_g = \frac{\Delta M}{M} = \frac{i_d F_L + B_O + \Delta B}{F_L + F_P + B_L + (1-R)D} = \frac{\frac{i_d F_L}{B} + \frac{B_O}{B} + p}{\frac{F_L}{B} + \frac{F_P}{B} + \frac{B_L}{B} + \frac{(1-R)D}{B}}$$

First, the extension of discount loans to the banks increases the excess reserves of banks and thus bank lending to the industrial sector. Federal Reserve loans (F_L) must, therefore, be added to the aggregate loan amount (or excess reserves) shown in the denominator of the equation. The banks lend reserves they acquire from discount loans and open market purchases by the central bank. They also lend their bank loan capital and the excess reserves they acquire from cash deposits.

Second, the discount loans are not provided at zero cost to the banks. The banks must pay the discount rate (i_d) for these loans. Therefore, the interest received on loans must be sufficient to cover the interest expense of discount loans. For that reason, i_dF_L is included in the numerator of the equation. This product represents the interest expense of the discount loans. If the commercial banks receive the interest implied in the numerator from the industrial sector, then they will be able to pay the interest expense of the discount loans and their operating expenses and still have enough remaining to equal the average profit.

These changes show how a change in the discount rate affects the general rate of interest and the market rate of interest. When the central bank increases the discount rate, the general rate of interest must rise. The cost of servicing discount loans increases, and the banks must pass this higher cost along to industry if they are to continue to receive the general rate of profit on bank capital. On the other hand, if the central bank reduces the discount rate, then the general rate of interest must fall because the fall in interest expenses allows the banks to earn the general rate of profit on bank capital even as it charges a lower rate of interest on its industrial loans.

Less obvious is the impact that discount lending has on the general rate of interest. Discount loans (F_L) are present in the denominator and the numerator. That is, the discount lending leads to more bank lending, which puts downward pressure on the general rate of interest. At the same time, it increases the interest expenses of the bank, which puts upward pressure on the general rate of interest. The question we must answer then is where the increase in discount loans will have the largest relative impact.

To answer this question, it helps to use a bit of calculus. Using the quotient rule from calculus, we can differentiate the general rate of interest with respect to Fed discount loans as follows:

$$\frac{di_g}{dF_L} = \frac{i_d(F_L+F_P+B_L+(1-R)D-(i_dF_L+B_O+\Delta B)}{(F_L+F_P+B_L+(1-R)D)^2}$$

If we set this derivative equal to zero, then we can determine what the discount rate must equal so that discount loans have zero impact on the general rate of interest.

$$\frac{di_g}{dF_L} = 0$$

If the fraction is to equal zero, then the numerator must equal zero. In other words:

$$\frac{di_g}{dF_L} = 0 \Leftrightarrow i_d(F_L + F_P + B_L + (1-R)D - (i_dF_L + B_O + \Delta B) = 0$$

Now we find the discount rate that solves the equation (i_d*):

$$i_d(F_L + F_P + B_L + (1-R)D) = i_dF_L + B_O + \Delta B$$
$$i_dF_L + i_dF_P + i_dB_L + i_d(1-R)D = i_dF_L + B_O + \Delta B$$
$$i_d(F_P + B_L + (1-R)D) = B_O + \Delta B$$
$$i_d^* = \frac{B_O+\Delta B}{F_P+B_L+(1-R)D}$$

This result is interesting. It shows that discount loans will have no impact on the general rate of interest if the discount rate equals the general rate of interest when no discount lending occurs. We will refer to i_d* as the **zero-impact discount rate** because it implies that discount loans will have no impact on the general rate of interest. Two additional results may also be stated:

$$\frac{di_g}{dF_L} > 0 \Leftrightarrow i_d > i_d^*$$

$$\frac{di_g}{dF_L} < 0 \Leftrightarrow i_d < i_d^*$$

That is, if the discount rate is relatively high (greater than i_d*), then increased discount lending will raise the general rate of interest due to the high interest expense of the discount loans. On the other hand, if the discount rate is relatively low (lower than i_d*), then increased discount lending will reduce the general rate of interest due to the low interest expense of the discount loans. Because the discount rate has been historically set at a low level, the impact of discount lending has tended to reduce interest rates, but it is theoretically possible that such lending could increase the general and market rates of interest.

The discount rate shown in Table 17.2 is below the level of the general rate of interest without any discount lending and only open market purchases (i_d* = 5.13%) as shown in Table 17.1. Therefore, the discount lending causes the general rate of interest to fall to 4.35%.

Table 17.3 shows the case of a relatively high discount rate of 6%.

Table 17.3: An Example with Central Bank Discount Lending (A High Discount Rate)

Industrial Sector		Financial Sector	
Industrial Rate of Profit	25.00%	Financial Rate of Profit	25.00%
Non-Borrowed Capital ($)	305,000	Bank Operating Capital ($)	5,000
Borrowed Capital (Loan Capital) ($)	390,000	Bank Loan Capital ($)	35,000
Total Productive Capital ($)	695,000	Total Deposits ($)	200,000
Total Profit ($)	173,750	Reserve Requirement Ratio	20.00%
Interest Paid to Banks ($)	20,850	**Total Bank Capital ($)**	40,000
Total Annual Social Product ($)	868,750	**General Rate of Interest (Prime Rate)**	≈5.35%
		Loan Interest Received or Gross Bank Profit ($)	20,850
Aggregate Profits ($)	183,750	**Net Bank Profit ($)**	10,000
Total Social Capital ($)	735,000	**Interest Paid to the Central Bank ($) (interest from discount loans)**	5,850
General Rate of Profit (Gross)	25.00%	Open Market Fed Purchases ($)	97,500
Interest Paid to Central Bank ($) (interest from government bonds)	5,212.50	Discount Loans	97,500
		Discount Rate	6.00%

The discount rate of 6% shown in Table 17.3 is above i_d^*. Therefore, when the discount lending occurs, the interest expense is so great that the general rate of interest rises to 5.35%.

Table 17.4 shows the case of a discount rate that is exactly equal to i_d^*.

Table 17.4: An Example with Central Bank Discount Lending ($i_d=i_d$*)

Industrial Sector		Financial Sector	
Industrial Rate of Profit	25.00%	Financial Rate of Profit	25.00%
Non-Borrowed Capital ($)	305,000	Bank Operating Capital ($)	5,000
Borrowed Capital (Loan Capital) ($)	390,000	Bank Loan Capital ($)	35,000
Total Productive Capital ($)	695,000	Total Deposits ($)	200,000
Total Profit ($)	173,750	Reserve Requirement Ratio	20.00%
Interest Paid to Banks ($)	20,000	**Total Bank Capital ($)**	40,000
Total Annual Social Product ($)	868,750	**General Rate of Interest (Prime Rate)**	≈5.13%
		Loan Interest Received or Gross Bank Profit ($)	20,000
Aggregate Profits ($)	183,750	**Net Bank Profit ($)**	10,000
Total Social Capital ($)	735,000	**Interest Paid to the Central Bank ($) (interest from discount loans)**	5,000
General Rate of Profit (Gross)	25.00%	Open Market Fed Purchases ($)	97,500
Interest Paid to Central Bank ($) (interest from government bonds)	5,000	Discount Loans	97,500
		Discount Rate	5.13%

When this situation arises, the general rate of interest does not change when the discount lending occurs. The economy expands even as the rate of interest remains the same.

The analysis in this section has explained how interest rates and the scale of economic activity changes when the central bank employs its monetary policy tools. The conclusions resemble those reached in neoclassical economic theory, but they are obtained using a Marxian framework in which industrial capitalists, banking capitalists, and the central bank share in the distribution of profits that the working class produces.

Following the Economic News[13]

As of February 2018, the Federal Reserve was expected to approach monetary policy, under the new Fed Chair Jerome Powell, in a manner resembling the approach of his predecessor Janet Yellen. Because of the rapid growth of the economy, gradual interest rate increases are expected to continue throughout 2018 as reported in *The Wall Street Journal*. The recent passage of a $1.5 trillion tax cut over the next ten years and the $300 billion federal spending increase over the next two years have also led to predictions of future Fed interest rate increases, according to WSJ writer Nick Timiraos. The concern that the U.S. economy might experience inflation due to these growth-promoting factors has led the Fed to pursue a contractionary monetary policy. Selling bonds to commercial banks or the public is the best strategy for removing reserves from the banking system, which will lead to less lending and higher interest rates. In neoclassical terms, the higher interest rates should discourage investment spending and net export spending because the higher interest rates make it costlier for firms to borrow and encourage an appreciation of the dollar, which makes U.S. exports relatively more expensive and imports relatively cheaper. The reduction in aggregate spending should reduce inflationary pressure in the economy. In Marxian terms, the Fed sales of securities are expected to reduce open market Fed purchases (F_p), which will raise the general rate of interest and the market rate of interest. It also will lead to a contraction of bank lending to the industrial sector and will reduce the total social product. Both frameworks suggest that the Fed's monetary contraction will raise interest rates and lead to a contraction of economic activity. Again, they arrive at these conclusions using entirely different theories.

Summary of Key Points

1. Monetary policy refers to the use of money

supply changes to influence aggregate production, the aggregate price level, and the level of unemployment.
2. The Federal Reserve System consists of a system of 12 Federal Reserve Banks, and the Federal Reserve Board of Governors and the Federal Open Market Committee (FOMC) determine monetary policy.
3. The Fed acts as a supervisor of banks, a lender of last resort, an issuer of currency, a banker for commercial banks, a processor of checks, a banker for the federal government, and a regulator of the money supply.
4. The Fed influences the money supply using open market purchases and sales of securities, discount lending, and adjustments to the reserve requirement ratio (R).
5. Expansionary monetary policy involves monetary expansions and interest rate reductions to boost investment spending and net export spending.
6. Contractionary monetary policy involves monetary contractions and interest rate increases to discourage investment spending and net export spending.
7. Expansionary monetary policy involves Fed purchases of bonds from banks, discount rate reductions, and reserve requirement ratio reductions.
8. Contractionary monetary policy involves Fed sales of bonds to banks, discount rate increases, and reserve requirement ratio increases.
9. The Quantity Equation is an identity that shows how the money supply, the velocity of money,

the aggregate price level, and the level of real output are all related.

10. Whereas the Quantity Theory of Money emphasizes the relative stability and predictability of the velocity of money, Post-Keynesian economists emphasize the instability and unpredictability of money velocity.
11. Post-Keynesian economists argue that the class conflict over the distribution of income leads to a wage-price inflationary spiral, which causes banks and the central bank to increase the money supply, thus providing an endogenous explanation for money supply growth.
12. The Marxian theory of fiat money treats the price of an individual commodity as a fraction of the effective money supply (MV) where fiat paper money represents a claim to other commodities in circulation.
13. Money wages and real wages have risen throughout U.S. economic history due to inflation and productivity growth. Nevertheless, American workers have experienced rising rates of exploitation due to the reduction in the necessary labor time required to produce the consumption bundle that workers consume and the increase in the surplus labor time performed during the workday.
14. When the central bank purchases securities from commercial banks, the increased lending to industrial capitalists pushes down the general and market rates of interest and leads to an economic expansion.
15. When the central bank increases discount lending to commercial banks, the increased

lending to industrial capitalists leads to an economic expansion. It also pushes down the general rate of interest so long as the discount rate is below the zero-impact discount rate.

16. When the central bank reduces the discount rate, the general and market rates of interest fall. When the central bank raises the discount rate, the general and market rates of interest rise.

List of Key Terms

Monetary policy

Federal Reserve

Federal Reserve Banks (FRBs)

Federal Reserve Board of Governors

Chair of the Federal Reserve Board

Vice Chair of the Federal Reserve Board

Federal Open Market Committee (FOMC)

Quasi-public institution

Central bank independence

Supervisor of banks

Lender of last resort

Issuer of currency

Federal Reserve Notes

Banker for commercial banks

Processor of checks

Banker for the federal government

Foreign currency-denominated assets

Regulator of the money supply

Open market operations

Discount loans (F_L)

Discount rate (i_d)

Reserve requirement ratio (R)

Expansionary (easy) monetary policy

Contractionary (tight) monetary policy

Boom-bust policy

Quantity Equation

Velocity of money

Quantity Theory of Money

Debt monetization

Exogenous variable

Endogenous variable

Exogenous money supply

Wage-price inflationary spiral

Endogenous money supply

Marxian monetary theory

Marxian theory of fiat money

Effective money supply (MV)

Real wage (q)

Necessary labor (NL)

SNALT required per unit of commodity (e)

Money wage (nominal wage)

Variable capital (v)

Surplus labor (SL)

Value of money (VOM)

Federal Reserve open market purchases of securities (F_P)

Zero-impact discount rate (i_d*)

Problems for Review

1. Suppose the Fed purchases $20,000 in bonds from the Valpo Bank. How will the Fed's balance sheet change? How will the Valpo Bank's balance sheet change? Show the changes on the Fed's balance sheet and the Valpo Bank's balance sheet. What is the impact on the money supply likely to be (e.g., positive, negative)?
2. Suppose the Fed sells $30,000 worth of bonds to the Valpo Bank. How will the Fed's balance sheet

change? How will the Valpo Bank's balance sheet change? Show the changes on the Fed's balance sheet and the Valpo Bank's balance sheet. What is the impact on the money supply likely to be (e.g., positive, negative)?

3. Suppose the Fed grants a $15,000 discount loan to the Gary Bank. How will the Fed's balance sheet change? How will the Gary Bank's balance sheet change? Show the changes on the Fed's balance sheet and the Valpo Bank's balance sheet. What is the impact on the money supply likely to be (e.g., positive, negative)?
4. Suppose the Gary Bank repays the Fed for a $12,000 discount loan (ignore the interest payment). How will the Fed's balance sheet change? How will the Gary Bank's balance sheet change? Show the changes on the Fed's balance sheet and the Valpo Bank's balance sheet. What is the impact on the money supply likely to be (e.g., positive, negative)?
5. Suppose the Fed reduces the reserve requirement ratio from 15% to 10%. If the South Bend Bank has checkable deposits of $200,000 and $30,000 in reserves, then how is it affected due to the reduction in the required reserve ratio? Show the changes to the South Bend Bank's balance sheet. What is the impact on the money supply likely to be (e.g., positive, negative)?
6. Suppose the Fed raises the reserve requirement ratio from 20% to 25%. If the South Bend Bank has checkable deposits of $200,000 and $40,000 in reserves, then how is it affected due to the increase in the required reserve ratio? Show the

changes to the South Bend Bank's balance sheet. What is the impact on the money supply likely to be (e.g., positive, negative)?

7. Suppose the Fed increases the discount rate. According to neoclassical theory, what is expected to happen to bank reserves, the money supply, interest rates, investment spending, the value of the U.S. dollar, and net export spending?
8. Suppose the money supply is $2.5 trillion, the velocity of money is 6, and the aggregate price level is 3. What is the level of real output? Use the Quantity Equation.
9. Suppose a wage-price deflationary spiral occurs. That is, prices fall and firms insist on cutting workers' wages. As demand drops, prices fall further. Using Post-Keynesian theory, predict what will happen to money demand, the money supply, interest rates, the supply and demand for loanable funds, the supply and demand for bonds, and bond prices.
10. Suppose the SNALT embodied in a mobile phone is 0.5 hours and the aggregate labor time embodied in all circulating commodities is 20 billion hours. If the money supply is $2 trillion and the velocity of money is 5, then what is the fiat money price of the mobile phone?
11. Consider Table 17.4. Suppose that the discount rate falls to 0.25%. How is the general rate of interest affected? Carry out the calculation. Compare the 0.25% discount rate to the zero-impact discount rate. Compare the newly calculated general rate of interest to the general rate of interest when no discount lending occurs

(i.e., 5.13%). Are your comparisons consistent with your expectations? Explain.

Notes

1. See Mishkin (2006), p. 339.
2. Chisholm and McCarty (1981), p. 207, state that the "purpose is not to earn money but to stabilize American banking and to decide and carry out monetary policy."
3. The monetary policy tools described in this section are discussed in virtually all neoclassical macroeconomics textbooks.
4. Mishkin (2006), p. 393-410, provides a helpful model of the supply and demand for bank reserves, which shows how the relationship between the discount rate and the federal funds rate influences the supply curve in the market for bank reserves.
5. In this discussion, it is assumed that only one interest rate exists in the economy. Many interest rates exist in the economy, but they tend to move together and so it is useful to refer to only one interest rate. See Mankiw (1997), p. 58, for a discussion of this point. The Fed has the most direct impact on the federal funds rate when it uses its policy tools, which then affects the general structure of interest rates throughout the economy.
6. Neoclassical textbooks recognize the increased instability of the M1 money velocity beginning in the early 1980s. For example, see OpenStax College (2014), p. 655.
7. A similar equation, based on one that Sidney Weintraub presented, is found in Snowdon, et al. (1994), p. 372.
8. Although I refer to the circulating paper money supply

throughout this discussion, the symbols may take electronic form and so checkable deposits may also be included here.

9. Saros (2002). I prove this result for the case of V = 1 in Saros (2006) and Saros (2007), p. 407-415. Fred Moseley has proven the result for the general case where V is equal to any value, which is a major contribution to Marxian monetary theory. Moseley conclusively proves that Marx's theory of value does not depend on the widespread use of a money commodity such as gold. See Moseley (2004) and Moseley (2011). I was unaware of Moseley's 2004 draft until after the RRPE review process ended for my 2006 submission, which was published in abbreviated form in 2007. I am thankful to Prof. Moseley for recognizing my development of these concepts as "completely independent" of his 2004 draft (see Moseley (2011), p. 102), although the development of my diagrammatic approach was certainly influenced by my reading of the introduction to his edited volume. See Moseley (2005).
10. See Hunt (2002), p. 244. Hunt explains that, despite what critics argue, nothing can be found in Marx's mature writings to suggest that he associated a falling real wage for workers with their increasing misery.
11. See Wolff and Resnick (2012), p. 192-195.
12. Moseley (2004/2011) has written the equation this way.
13. Timiraos, Nick. "For the Federal Reserve, a New Chief but Same Interest-Rate Path." *The Wall Street Journal*. Web. Updated February 16, 2018. Accessed on February 18, 2018.

Chapter 18

THEORIES OF GOVERNMENT BUDGET DEFICITS AND DEBT

Goals and Objectives:

In this chapter, we will do the following:

1. *Explore* the components of the federal budget and the federal debt
2. *Analyze* the macroeconomic impact of expansionary and contractionary fiscal policy
3. *Examine* the macroeconomic impact of government budget deficits and surpluses
4. *Distinguish* between marginal tax rates and average tax rates in different taxation systems
5. *Develop* the concept of a tax rate multiplier
6. *Investigate* the implications of fiscal policy for macroeconomic stability
7. *Demonstrate* the value of a concept referred to as the full employment budget
8. *Contrast* Keynesian full employment policies with neoclassical austerity policies
9. *Evaluate* a Marxian theory of government borrowing and debt

In Chapter 17, we investigated the role of the central bank and how monetary policy may be used to influence aggregate output, employment, and the general price level from several different theoretical perspectives. In this chapter, we consider the role of the federal government and how fiscal policy may be used to influence key macroeconomic variables from different vantage points, including neoclassical, Keynesian, and Marxian perspectives. To set the stage, we will define such concepts as government deficits and government debt and then look at the structure of the federal budget for Fiscal Year 2015. Next, we will investigate the macroeconomic impacts of expansionary and contractionary fiscal policy and government deficits and surpluses from neoclassical and Keynesian perspectives. We will then look at different systems of taxation and how to develop a tax rate multiplier to be used in the Keynesian Cross model. Additional topics include the implications of fiscal policy for macroeconomic stability, the usefulness of the concept of the full employment budget, and the contrast between Keynesian full employment policies and neoclassical austerity policies. We will conclude with a Marxian analysis of government borrowing and government debt.

The Federal Budget and the Federal Debt

Fiscal policy pertains to the use of government spending and taxation to influence aggregate output, employment and the price level. The federal government spends a great deal of money, but it also receives a great deal of money through tax collections and other sources. When government outlays and receipts for the year do not match, we say that the budget is out of balance. When government outlays and government receipts for the year do match, then we refer to a **balanced budget**.

When government outlays exceed government receipts for the year, then we say that a **budget deficit** exists. Finally, when government outlays fall short of government receipts for the year, then we say that a **budget surplus** exists. Using R to represent government receipts and O to represent government outlays, we can list the possibilities as follows:

$O = R \Leftrightarrow A\ Balanced\ Budget$

$O > R \Leftrightarrow A\ Budget\ Deficit$

$O < R \Leftrightarrow A\ Budget\ Surplus$

Although we speak in terms of an annual budget, it is not the calendar year that we have in mind but rather the government's fiscal year, which begins on October 1 of each year and ends the following September 30. For example, Fiscal Year 2018 (FY2018) includes the date October 5, 2017 but not September 25, 2017.

To develop a sense of which items the federal budget includes, let's consider the figures for FY2015. Table 18.1 shows the Unified Federal Budget for FY2015, which lists all Federal outlays and receipts for that year.[1]

Table 18.1: The FY2015 Federal Budget, in millions of dollars

Receipts (+)		Outlays (-)	
On-Budget		**On-Budget**	
Individual Income Taxes	1,540,802	National Defense	(589,659)
Corporation Income Taxes	343,797	Human Resources	(1,850,064)
Social Insurance and Retirement Receipts	294,885	Physical Resources	(116,879)
Excise Taxes	98,279	Net Interest	(319,149)
Estate and Gift Taxes	19,232	Other	(169,350)
Customs Duties and Fees	35,041	Undistributed Offsetting Receipts	99,795
Federal Reserve Deposits	96,468	**On-Budget Outlays**	(2,945,306)
Other	51,011	**Off-Budget**	
On-Budget Receipts	2,479,515	Postal Service	1,710
Off-Budget		Social Security	(856,763)
Social Insurance and Retirement Receipts	770,372	Interest received by on-budget trust funds	95,968
Off-Budget Receipts	770,372	Undistributed offsetting receipts	16,008
Total Receipts	3,249,887	**Off-Budget Outlays**	(743,077)
UNIFIED BUDGET DEFICIT	(438,496)	**Total Outlays**	(3,688,383)

Only some items have been listed in Table 18.1 with the remaining items included in "Other" categories. Also, the receipts and outlays have been divided according to whether they are "on-budget" or "off-budget." The federal government's unified budget includes all receipts and outlays of the federal government. Some receipts and outlays, however, are treated as off-budget because of a desire to protect them from lawmakers, in the case of Social Security, for example, or because it is viewed as a part of the budget that should be able to achieve balance independently, as in the case of the U.S. Postal Service.

To calculate the **unified budget deficit** or **unified budget surplus**, we simply add up the total receipts and subtract the total outlays to obtain a unified budget deficit of $438.496 billion. It is also possible to calculate the **on-budget deficit** or **on-budget surplus** and the **off-budget deficit** or **off-budget surplus**. In the former case, we take the on-budget receipts and subtract the on-budget outlays to obtain an on-budget deficit of $465.791 billion. In the latter case, we take the off-budget receipts and subtract the off-budget outlays to obtain an off-budget surplus of $27.295 billion.

The on-budget deficit and the off-budget surplus may be added together to obtain the unified budget deficit. To understand why, consider the following equation:

$$Unified\ Budget\ Balance = R - O = (R_N + R_F) - (O_N + O_F)$$

In the above equation, the sum of on-budget receipts (R_N) and off-budget receipts (R_F) is calculated and then we subtract the sum of on-budget outlays (O_N) and off-budget outlays (O_F). Rearranging the terms, we obtain the following result:

$$Unified\ Budget\ Balance = (R_N - O_N) + (R_F - O_F) = on\ budget\ balance + off\ budget\ balance$$

That is, we can simply sum together the on-budget balance and the off-budget balance to obtain the unified budget balance. In this example, if we add the off-budget surplus to the on-budget deficit, then we obtain the unified budget deficit of $438.496 billion. This example demonstrates the benefit of reporting the off-budget balance separately from the on-budget balance. Because the Social Security program has had many years of surpluses, its inclusion in the unified budget has made the federal deficit appear smaller than it is. If we focus on the part of the budget over which lawmakers have more control from year to year (i.e., the on-budget deficit), then we can see that the federal deficit was somewhat larger in FY2015 and has been far larger than the unified budget deficit in some years due to large Social Security surpluses.

The complete federal budget contains many items, but federal outlays can be divided into three major categories. Frequently, the **appropriated programs** or **discretionary programs** are grouped together because they require Congress to pass annual appropriations bills. These outlays include spending on agriculture, defense, education, energy, homeland security, health and human services, housing and urban development, environmental protection, and so on. **Mandatory spending**, on other hand, includes **entitlement programs** that do not depend on Congress to pass annual appropriations bills and where the outlays depend on who qualifies for benefits under federal law. Examples include Social Security, Medicare, and Medicaid. The final major component is **net interest**, which accounts for interest paid to owners of U.S. government bonds less the interest that the government receives. The total outlays for FY2015 amounted to $3,688.383 billion, or $3.688383 trillion, as shown in Table 18.1.

Federal receipts for FY2015 include several different sources, including individual income taxes, corporate income taxes, Social Security payroll taxes, Medicare payroll taxes, unemployment insurance taxes, excise taxes, estate and gift taxes, customs duties, and profit distributions from the Federal Reserve. The total receipts for FY2015 amounted to $3,249.887 billion, or $3.249887 trillion, as shown in Table 18.1.

Table 18.2 shows federal receipts and federal outlays as percentages of their totals for FY2015.[2]

Table 18.2: The FY2015 Federal Budget, percentages

Receipts (+) as a Percentage of Total Receipts		Outlays (-) as a Percentage of Total Outlays	
On-Budget		**On-Budget**	
Individual Income Taxes	47.41%	National Defense	15.99%
Corporation Income Taxes	10.58%	Human Resources	50.16%
Social Insurance and Retirement Receipts	9.07%	Physical Resources	3.17%
Excise Taxes	3.02%	Net Interest	8.65%
Estate and Gift Taxes	0.59%	Other	4.59%
Customs Duties and Fees	1.08%	Undistributed Offsetting Receipts	-2.71%
Federal Reserve Deposits	2.97%	**On-Budget Outlays**	79.85%
Other	1.57%	**Off-Budget**	
On-Budget Receipts	**76.30%**	Postal Service	-0.05%
Off-Budget		Social Security	23.23%
Social Insurance and Retirement Receipts	23.70%	Interest received by on-budget trust funds	-2.60%
Off-Budget Receipts	**23.70%**	Undistributed offsetting receipts	-0.43%
Total Receipts	**100.00%**	**Off-Budget Outlays**	20.15%
UNIFIED BUDGET DEFICIT	--	**Total Outlays**	100.00%

The percentages reveal which items represent the largest shares of the total receipts and total outlays. In terms of receipts, individual incomes taxes represent the largest share, followed by Social Security receipts and corporate income taxes. In terms of outlays, Human Resources, Social Security, and National Defense represent the largest shares, followed by net interest payments on the debt.

The **federal debt** represents the entire accumulated debt of the federal government minus whatever has been repaid over the years. In any given year, when federal receipts fall short of federal outlays, it is necessary for the federal government to borrow to make up the difference. These borrowings add to the federal debt. Therefore, the federal debt may be thought of as the accumulation of past federal deficits (less any repayments out of federal surpluses). Deficits and surpluses, therefore, represent flow variables because they are measured on an annual basis. The national debt, on the other hand, represents a stock variable because we can identify the total amount owed at a given point in time.

At the end of FY2015, total gross federal debt amounted to $18.120 trillion. Most of the borrowed funds (about $18.094 trillion) was acquired through the issuance and sale of Treasury securities. The remainder (about $26 billion) was acquired via the issuance and sale of federal agency bonds. This information is presented in Table 18.3.

Table 18.3: The Federal Debt at the end of 2015, in billions of dollars

Federal Government Financing and Debt, in billions of dollars	
Gross Federal Debt issued by:	**18,120**
Debt issued by Treasury	18,094
Debt issued by other agencies	26
Gross Federal Debt held by:	**18,120**
Debt held by Government Accounts	5,003
Debt held by Federal Reserve Banks	2,461.9
Debt held by the rest of the public	10,654.8

Source: Office of Management and Budget. Fiscal Year 2017. 'Budget of the U.S. Government. Table S-13. Federal Government Financing and Debt. U.S. Government Printing Office. Washington. 2016. Pp. 165. Note: Some rounding error exists in the calculation of these totals.

The buyers of these government bonds are the holders of U.S. government debt obligations. These bonds represent promises of the U.S. government to repay the face values and to maintain regular interest payments until the bonds mature. Who owns these bonds? Interestingly, a large portion of these bonds (about $5.003 trillion) is held in

U.S. Government accounts such as the Social Security Trust Fund. When the Social Security program has a surplus, for example, this amount is invested in a special category of U.S. Treasury bonds. The Federal Reserve Banks hold another portion of the debt (about $2.4619 trillion). The remaining $10.6548 trillion is held by the rest of the public. It is surprising to many people that such a large percentage of the federal debt (about 41.2%) is owed to the federal government or to the nation's quasi-public central bank.

Domestic and foreign investors own the part of the debt that the federal government and the Fed do not own. At the end of December 2015, the total foreign holdings of Treasury securities amounted to $6,146.2 billion, or $6.1462 trillion. China and Japan are the largest holders of U.S. Treasury securities with more than $1.1 trillion each in FY2015. All foreign nations with more than $100 billion of Treasury securities in FY2015 are listed in Table 18.4.

Table 18.4: Major Foreign Holders of Treasury Securities in December 2015, billions of dollars

Country	Holdings at End of Period	Country	Holdings at End of Period
China	1,246.1	Hong Kong	200.1
Japan	1,122.4	Luxembourg	199.6
Ireland	264.4	Taiwan	178.7
Brazil	254.8	Belgium	121.7
Cayman Islands	249.8	Saudi Arabia	118.9
Switzerland	231.7	India	116.8
United Kingdom	207.1	Singapore	110.3
All Other	1,523.9	**TOTAL**	**6,146.2**

Source: U.S. Treasury. "Major Holders of Treasury Securities." MFH tables, History back to March 2000. Web. Accessed on February 22, 2018. https://www.treasury.gov/resource-center/data-chart-center/tic/Pages/ticsec2.aspx.
Note: Some rounding error exists in the calculation of the total.

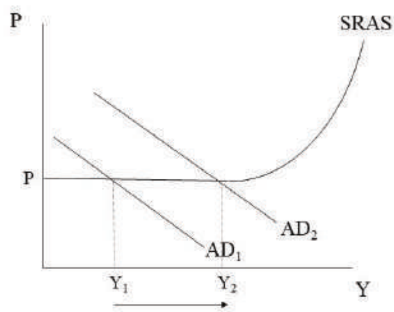
The foreign holdings of Treasury securities amount to approximately 1/3 of the federal debt. These basic facts about the federal budget and the federal debt will be useful as we consider the macroeconomic impacts of each from a variety of perspectives.

Expansionary Fiscal Policy versus Contractionary Fiscal Policy

The federal government may use fiscal policy to pursue an economic expansion. In that case, it implements an **expansionary fiscal policy** with the aim of promoting higher aggregate output and employment. On the other hand, the federal government might use fiscal policy to restrict the growth of output and employment. The government is then said to be implementing a **contractionary fiscal policy**, which it might implement due to its fear of inflation.

Expansionary fiscal policy might be implemented in several different ways. To pursue an economic expansion, the federal government might cut taxes, increase government spending, or combine the two policies.[3] Each measure will stimulate aggregate demand, and if implemented in the short run when prices are sticky, the rise in aggregate demand will increase real output and employment as shown in Figure 18.1.

Figure 18.1: The Impact of Expansionary Fiscal Policy in the Short Run – Lump Sum Tax Cuts and/or Government Spending Increases



The precise impact on real output, however, depends on the size of the lump sum tax multiplier and the government expenditures multiplier.

The reader should recall that the lump sum tax multiplier is the following:

$$\frac{\Delta Y}{\Delta T} = \frac{-mpc}{1-mpc}$$

Similarly, the government expenditures multiplier is the following:

$$\frac{\Delta Y}{\Delta G} = \frac{1}{1-mpc}$$

Suppose that the government cuts taxes by $50 billion, and the marginal propensity to consume (mpc) is 3/4. Then the lump sum tax multiplier is equal to -3 and we can calculate the change in real output as follows:

$$\Delta Y = \frac{-mpc}{1-mpc} \cdot \Delta T = \frac{-\frac{3}{4}}{1-\frac{3}{4}} \cdot (-50B) = (-3) \cdot (-50B) = +\$150B$$

That is, a tax cut of $50 billion increases real output by $150 billion. The reason for the multiplier is that households have more after-tax income to spend, which triggers additional rounds of household spending. Ultimately, real output increases three times as much as the tax cut.

Now suppose that the government increases spending by $50 billion, and the mpc is ¾. The government expenditures multiplier in this case is equal to 4, and we can calculate the change in real output as follows:

$$\Delta Y = \frac{1}{1-mpc} \cdot \Delta G = \frac{1}{1-\frac{3}{4}} \cdot (50B) = (4) \cdot (50B) = +\$200B$$

A government spending increase of $50 billion increases real output by $200 billion. The reason for the multiplier is that households receive the government spending as income and then spend it, which triggers even more consumption. In the end, real output rises by four times the initial increase in spending.

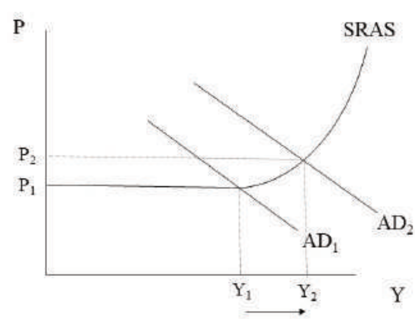
Finally, consider a combination of government spending increases and tax cuts, not unlike the policy pursued in early 2009 when Congress passed the American Recovery Act. This piece of legislation included a mix of tax cuts and spending increases aimed at boosting economic activity during the Great Recession. Suppose that the government increases spending by $20 billion and cuts taxes by $40 billion. Continue to assume that the mpc is ¾. In this case, real output receives a boost from both sources as follows:

$$\Delta Y = \frac{1}{1-mpc} \cdot \Delta G + \frac{-mpc}{1-mpc} \cdot \Delta T = (4) \cdot (20B) + (-3)(-40B) = +\$200B$$

This example shows that a combination of a tax cut and a smaller spending increase can achieve the same increase in real output as the larger spending increase. In all three cases, aggregate demand experiences a rightward shift, and real output and employment expand as a result.

It is also worth noting that the full multiplier effects are felt only if prices are sticky in the short run as shown in Figure 18.1. The situation is somewhat different if prices are at least partly flexible as shown in Figure 18.2.

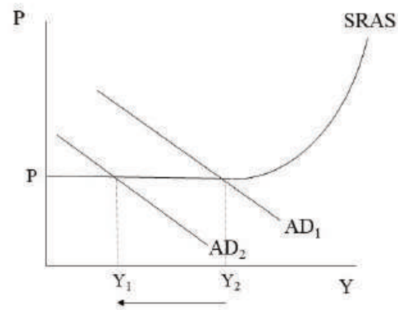
Figure 18.2: The Impact of Expansionary Fiscal Policy with Some Price Flexibility – Lump Sum Tax Cuts and/or Government Spending Increases



In Figure 18.2, the rise in real output is somewhat offset due to the rising price level. As the economy approaches the full employment level of real output, some of the aggregate demand increase leads to higher prices in addition to higher real output. Since the full impact of the rise in aggregate demand is not felt on real output, the multipliers do not fully function. In the extreme case where AD rises and intersects the vertical portion of the AS curve, prices are completely flexible and the multipliers are not operative at all with real output stuck at the full employment level. The case of perfect price flexibility is the neoclassical assumption whereas the assumption of sticky prices is the Keynesian assumption.

Contractionary fiscal policy might also be implemented in several different ways. To pursue an economic contraction, the federal government might raise taxes, cut government spending, or combine the two policies. Each measure will reduce aggregate demand, and if implemented in the short run when prices are sticky, the drop in aggregate demand will decrease real output and employment as shown in Figure 18.3.

Figure 18.3: The Impact of Contractionary Fiscal Policy in the Short Run – Lump Sum Tax Increases and/or Government Spending Reductions



As before, the precise impact on real output depends on the size of the lump sum tax multiplier and the government expenditures multiplier.

Suppose that the government raises taxes by $50 billion, and the mpc is 3/4. Then the lump sum tax multiplier is equal to -3 and we can calculate the change in real output as follows:

$$\Delta Y = \frac{-mpc}{1-mpc} \cdot \Delta T = \frac{-\frac{3}{4}}{1-\frac{3}{4}} \cdot (50B) = (-3) \cdot (50B) = -\$150B$$

That is, a tax increase of $50 billion reduces real output by $150 billion. The reason for the multiplier effect here is that households have less after-tax income to spend, which triggers additional reductions of household spending. Ultimately, real output falls three times as much as the tax increase.

Now suppose that the government reduces spending by $50 billion, and the mpc is ¾. The government

expenditures multiplier in this case is equal to 4, and we can calculate the change in real output as follows:
$\Delta Y = \frac{1}{1-mpc} \cdot \Delta G = \frac{1}{1-\frac{3}{4}} \cdot (-50B) = (4) \cdot (-50B) = -\$200B$

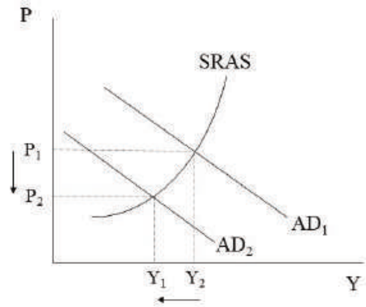
A government spending reduction of $50 billion decreases real output by $200 billion. The reason for the multiplier effect in this case is that households no longer receive the government spending as income and so cannot spend it, which triggers even less consumption. In the end, real output falls by four times the initial reduction in spending.

Finally, consider a combination of government spending reductions and tax increases. Suppose that the government reduces spending by $20 billion and raises taxes by $40 billion. In this case, real output receives a boost from both sources as follows:
$\Delta Y = \frac{1}{1-mpc} \cdot \Delta G + \frac{-mpc}{1-mpc} \cdot \Delta T = (4) \cdot (-20B) + (-3)(40B) = -\$200B$

This example shows that a combination of a tax increase and a smaller spending reduction can achieve the same decrease in real output as the larger spending reduction. In all three cases, aggregate demand experiences a leftward shift and real output and employment contract as a result. The benefit to the economy is that less danger exists that a sudden rise in aggregate demand will lead to inflation.

It is also worth noting that the full multiplier effects are felt only if prices are sticky in the short run and downwardly inflexible as shown in Figure 18.3. The situation is somewhat different if prices are at least partly flexible as shown in Figure 18.4.

Figure 18.4: The Impact of Contractionary Fiscal Policy with Some Price Flexibility – Lump Sum Tax Increases and/or Government Spending Reductions



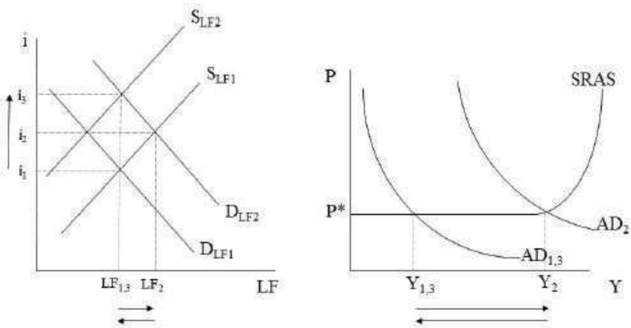
In Figure 18.4, the fall in real output is somewhat offset due to the falling price level. As the economy contracts, some of the aggregate demand reduction leads to lower prices in addition to lower real output. Since the full impact of the fall in aggregate demand is not felt on real output, the multipliers do not fully function, and deflation occurs. In the extreme case where AD falls and intersects the vertical portion of the AS curve, prices are completely flexible and the multipliers are not operative at all with real output stuck at the full employment level even as AD and the price level fall.

The Macroeconomic Impacts of Government Budget Deficits and Surpluses

When the federal government runs a budget deficit, it typically borrows to make up the difference. Printing the money is another option, but modern economies have rejected this solution given its tendency to produce hyperinflation. When a government wishes to borrow to cover its budget deficit, it must issue bonds and sell them to the public. This entrance into the loanable funds market and the bond market has the potential to alter interest rates and bond prices, which then carries consequences for the rest of the economy.

Suppose the government is borrowing in the loanable funds market, which permits it to run a budget deficit. At the same time, assume that the central bank is tightly controlling the supply of loanable funds such that it always adjusts the supply of loanable funds to offset any change in the equilibrium quantity exchanged in this market. The increased government borrowing will raise the demand for loanable funds, which creates a shortage and drives up the interest rate. The equilibrium quantity exchanged rises too. Because the central bank is regulating the quantity of loanable funds, it reduces the supply of loanable funds using its monetary policy tools. These changes to the supply and demand for loanable funds are shown in Figure 18.5.

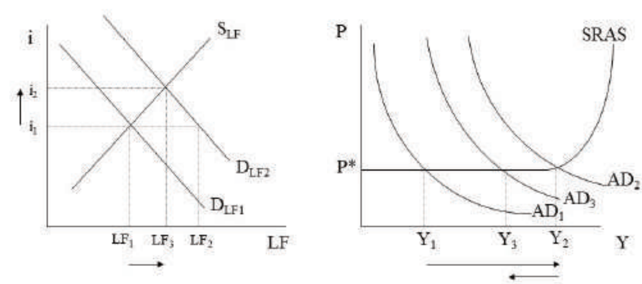
Figure 18.5: Complete Crowding Out of Private Investment due to Government Borrowing



The reduction in the supply of loanable funds causes a leftward shift of the supply curve, which drives the interest rate up further but causes the equilibrium quantity exchanged to return to its original level. The refusal of the central bank to accommodate the increase in government borrowing has led to a higher interest rate, which completely crowds out debt-financed private investment and consumption. That is, government spending on goods and services expands entirely at the expense of private sector spending on goods and services. In terms of the AD/AS model, the increase in debt-financed government spending will increase aggregate demand, but the higher interest rates reduce investment spending and consumer spending, which leaves AD unchanged in the end as shown in Figure 18.5. In this scenario involving **complete crowding out** of private investment, the *composition* of total output changes with a shift away from investment spending and towards government spending, but the *aggregate* output is not affected.[4] It is also worth noting that the multiplier effect is completely inoperative in this situation even though the price level is constant.

Another situation that might arise is one in which the central bank does not alter the supply of loanable funds in response to the increased demand for loanable funds when the government runs a deficit. In this situation, the demand curve shifts rightward, but the supply curve does not shift in the loanable funds market. This situation is depicted in Figure 18.6.

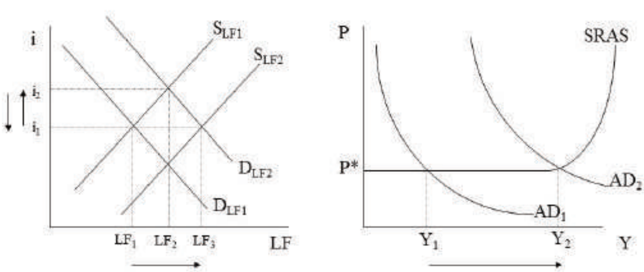
Figure 18.6: Partial Crowding Out of Private Investment due to Government Borrowing



The increased demand for loanable funds in Figure 18.6 causes the interest rate and the quantity exchanged to increase. Although the interest rate rises, it does not rise as much as it does when the central bank also restricts the supply of loanable funds. This scenario involves the **partial crowding out** of private investment spending and private consumer spending because the higher demand does manage to increase the equilibrium quantity in this market even as the interest rate rises. The partial crowding out is reflected in the movement along the new demand curve as the interest rate rises. In this case, the increase in government spending shifts the AD curve to the right, which then partially shifts back to the left due to the reduction in consumer spending and investment spending as the interest rate rises. The multiplier effect is lessened in this situation as well due to the rise in interest rates even though the price level is constant.

A final situation of interest is one in which the central bank completely accommodates the increased demand for loanable funds that occurs when the federal government runs a deficit. This situation of **complete accommodation** is represented in Figure 18.7.

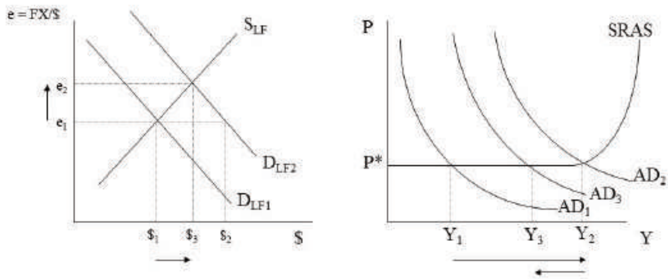
Figure 18.7: Zero Crowding Out of Private Investment due to Government Borrowing



In Figure 18.7 the demand for loanable funds increases due to government borrowing, but then the supply of loanable funds also rises as the central bank accommodates the increased demand for loans. The rightward shift of the demand curve raises the equilibrium quantity and the interest rate. The rightward shift of the supply further raises the equilibrium quantity but brings the interest rate back down to the original level. Because the interest rate does not change in this situation, no crowding out of private investment or private consumption occurs, and the full multiplier effect is operative if the price level is constant. This scenario was discussed in Chapter 17 in the context of Post-Keynesian endogenous money supply theory. The Fed accommodates the higher demand for loanable funds with an increase in the money supply. If prices are sticky, then the full impact of the higher money supply is felt on real output. If the rightward shift of AD occurred in the upward sloping portion of the SRAS curve, then the rise in the price level would lessen the impact on real output. Even with a rise in the price level, however, no crowding out occurs because the interest rate does not rise.

In addition to a possible reduction in investment spending and consumer spending following an increase in government deficit spending, it turns out that net export spending may also be affected. To understand the reason, consider what happens in a situation involving the partial crowding out of private investment. In that case, the government runs a deficit, which boosts the demand for loanable funds. The interest rate and the equilibrium quantity of loanable funds increase. The rise in the interest rate means that foreign investors will want to increase lending to the United States. To purchase U.S. dollar-denominated assets, these foreign investors must purchase U.S. dollars in the foreign exchange market. The higher demand for U.S. dollars will cause the foreign exchange value of the U.S. dollar to rise in a system of **flexible exchange rates**. This appreciation of the U.S. dollar is shown in Figure 18.8.

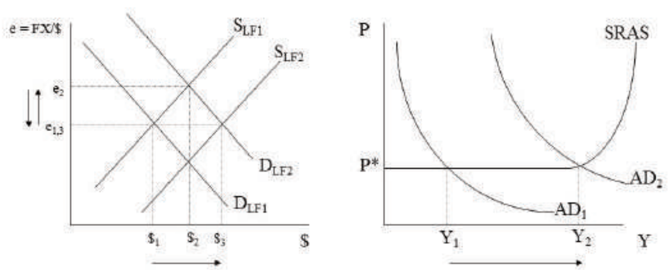
Figure 18.8: Partial Crowding Out due to Deficit Spending with Flexible Exchange Rates



When the dollar appreciates, U.S. exports become more expensive for foreign buyers and imports become cheaper for American buyers. The consequence is a reduction in U.S. exports and a rise in U.S. imports. Net exports thus fall, which causes a leftward shift of the AD curve, partially offsetting the rightward shift of AD due to higher government spending.

It is possible, however, that the central bank might act to offset the impact on the exchange rate. For example, when interest rates rise and the dollar appreciates, the central bank might try to stabilize the foreign exchange value of the dollar by selling U.S. dollars in the foreign exchange market as would be the case in a system of **fixed exchange rates**. The consequence will be an increase in the supply of U.S. dollars and a reduction of its value as shown in Figure 18.9.

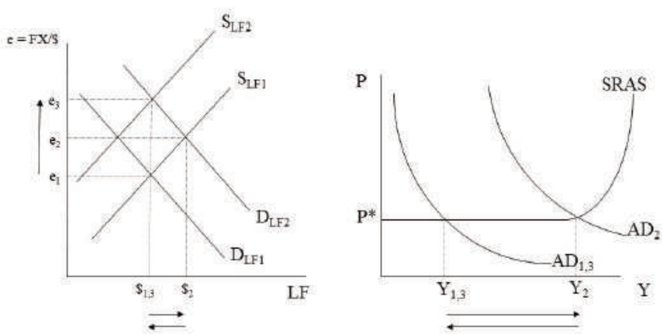
Figure 18.9: Partial Crowding Out due to Deficit Spending with Fixed Exchange Rates



If the exchange rate returns to its original level, then net exports should not be negatively affected even though the government deficit spending will still lead to a rise in interest rates and partial crowding out. The partial crowding out in this case would involve a leftward shift of AD from AD2 (not shown in the graph), but the shift would not be as large as that shown in Figure 18.8 because a reduction in net exports does not contribute to the decline in aggregate demand in this case.

A final possibility is that the central bank is more concerned about regulating the volume of dollar trading in the foreign exchange market. In this scenario, it acts to stabilize the equilibrium quantity exchanged of U.S. dollars. Suppose that the higher interest rates lead to a higher demand for U.S. dollars and so the central bank reduces the supply of U.S. dollars in the foreign exchange market, thereby shifting the supply curve to the left. This possibility is represented in Figure 18.10.

Figure 18.10: Partial Crowding due to Deficit Spending with a Regulated Trading Volume

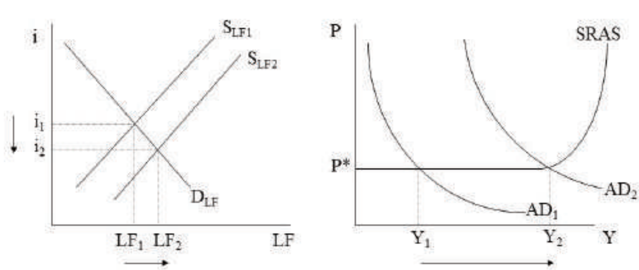


The consequence of the central bank's supply reduction is a further appreciation of the U.S. dollar even as it stabilizes the quantity exchanged of dollars. Figure 18.10 assumes that the crowding out of private investment and the negative impact on net exports exactly offset the expansionary impact of the deficit spending. The impact on U.S. net exports could be so extreme, however, that combined with the partial crowding out of private investment, the consequent drop in real output and employment could more than cancel the expansionary impact of the government deficit spending.

The examples in this section demonstrate that expansionary fiscal policy may be partly offset, completely offset, or not at all offset by action that the central bank takes. Coordination among policymakers is essential so that fiscal policy and monetary policy do not work against one another in terms of their impact on output, employment, and the price level.

A very different possibility is that the federal government manages to run a budget surplus. In that situation, it must decide how to use the surplus, and the repayment of federal debt is one possibility. If the government chooses this option, then it repays the bondholders. The bondholders then possess loanable funds that they will probably wish to lend again. The likely result then is an increase in the supply of loanable funds as shown in Figure 18.11.

Figure 18.11: The Macroeconomic Implications of Budget Surpluses and the Repayment of Federal Debt



In this example, the increased supply creates a surplus of loanable funds and pushes the rate of interest down. The drop in the rate of interest stimulates investment spending and consumer spending. Aggregate demand thus shifts to the right. If prices are sticky, the impact is an unambiguous gain for the economy with rising real output and employment and a stable price level. If the economy is close to full employment, however, the risk of the rise in the supply of loanable funds is demand-pull inflation.[5]

Marginal Tax Rates and Average Tax Rates

Up to this point, it has been assumed that the government simply appropriates tax revenue in one lump sum amount from the households each year. The lump sum tax (T) is not very realistic because households are taxed a specific percentage of their incomes. In fact, different tax rates apply to different income levels. For example, in 2017 a single taxpayer paid 10% of the first $9,325 of income in taxes to the federal government. For earnings between $9,325 and $37,950, a single taxpayer paid a 15% tax rate while still paying a 10% tax rate on the first $9,325 earned. The tax rates that apply to different income tax brackets are called marginal tax rates. **Marginal tax rates** tell us how much of an additional dollar of income is paid in taxes. **Income tax brackets** tell us the range of income to which a specific marginal tax rate applies. Table 18.5 shows the marginal tax rates and income tax brackets for a single taxpayer filing in 2017.

Table 18.5: Income Tax Brackets and Rates for a Single Filer, 2017

Income Tax Brackets and Rates for a Single Filer, 2017					
Marginal Tax Rate	Lowest Income in Bracket ($)	Highest Income in Bracket ($)	Median Income in Bracket ($)	Taxes Owed on the Median Income in the Bracket ($)	Average Tax Rate on the Median Income in the Bracket
10%	0	9,325	4,662.50	466.25	10.00%
15%	9,325	37,950	23,637.50	3,079.38	13.03%
25%	37,950	91,900	64,925	11,970	18.44%
28%	91,900	191,650	141,775	32,678.75	23.05%
33%	191,650	416,700	304,175	83,777	27.54%
35%	416,700	418,400	417,550	121,207.75	29.03%
39.6%	418,400	No upper limit	NA	NA	NA

Source: Tax Foundation. "2017 Tax Brackets." Tax Foundation. November 10, 2016. Web. Accessed on February 25, 2018. The author calculated the figures in the columns related to median income levels.

Table 18.5 also shows us how much tax is paid if one earns the median income in each tax bracket. The **median income** is the income level that is exactly halfway between the highest and lowest income levels in the tax bracket. For example, the median income levels in the 10% and 15% income tax brackets are equal to $4,662.50 and $23,637.50, respectively, and are calculated as follows:

$$Median\ Income\ for\ 10\%\ Income\ Tax\ Bracket = \frac{9{,}325+0}{2} = \$4,662.50$$

$$Median\ Income\ for\ 15\%\ Income\ Tax\ Bracket = \frac{37{,}950+9{,}325}{2} = \$23,637.50$$

To determine the taxes owed on the median income in any tax bracket, we cannot simply multiply that median income level by the marginal tax rate that applies to that tax bracket. To calculate the taxes owed in this manner would assume that a single tax rate applies to the entire income. The correct calculation requires that we multiply each increment of income up to that income level by the appropriate marginal income tax rate. For example, to calculate the taxes owed for someone who earns \$141,775 (the median income in the 28% income tax bracket), we do the following:

$Taxes\ owed\ on\ \$141,775 = (141,775-91,900)(28\%)+(91,900-37,950)(25\%)+(37,950-9,325)(15\%)+(9,325-0)(10\%) = \$32,678.75$

The taxes owed for other income levels are determined in a similar fashion. It is simple to determine the percentage of income that is owed in taxes, which is called the average tax rate. To calculate the **average tax rate**, simply divide the total taxes owed by the income level as follows:

$Average\ tax\ rate = \frac{Taxes\ Owed}{Income}$

For example, the median income level in the 33% income tax bracket is \$304,175, and the taxes owed are \$83,777. The average tax rate is calculated as follows:

$Average\ tax\ rate = \frac{83{,}777}{304{,}175} = 27.54\%$

The reader should notice that the marginal income tax rates increase with income level. As a result, the taxes owed rise more quickly than income as income increases. The result is a rising average tax rate, which is also visible in the table. When the average tax rate increases with income, the system of taxation is referred to as a **progressive tax system**. When a person earns more income in such a tax system, the total taxes owed increase but also the percentage of income paid in taxes increases. The justification for a tax system of this kind is that it makes possible the government redistribution of income and thus has the potential to reduce after-tax income inequality.

Those who oppose progressive taxation systems frequently advocate a single marginal tax rate that applies to all levels of income. In this case, the taxes owed and the income level rise at the same rate, which leaves the average tax rate unchanged. When the average tax rate remains constant at all levels of income, the system of taxation is referred to as a **proportional tax system**. A proportional tax is more commonly called a **flat tax**.

A third type of taxation system involves marginal tax rates that decrease as the income level rises. In this type of taxation system, the taxes owed increase more slowly than income rises. The result is a decline in the average tax rate as income rises. When the average tax rate falls as income rises, the taxation system is referred to as a **regressive tax system**.

Introducing a Flat Tax into the Consumption Function and the Keynesian Cross Model

Since the flat tax is the simplest system of taxation, with a single tax rate that applies to all income levels, let's consider how the introduction of a flat tax rate modifies the consumption function that we introduced in Chapter 13. A flat tax rate (t) is calculated as the ratio of taxes (T) to total income, which in this case is the level of real GDP.

$$t = \frac{T}{Y} \Rightarrow T = tY$$

That is, the total taxes owed are simply a fraction (t) of the total income earned. In Chapter 13, we wrote the after-tax consumption function as follows:

$$C_a = C_0 + mpc \cdot (Y - T)$$

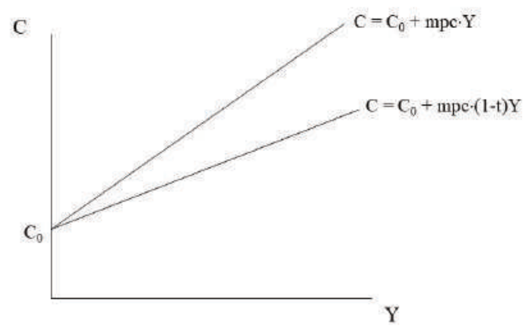
We can include the fact that the taxes owed are a function of Y as follows:

$$C_a = C_0 + mpc \cdot (Y - tY)$$

$$C_a = C_0 + mpc \cdot (1 - t)Y$$

This result shows that the flat tax has made the graph of the consumption function flatter. That is, the slope of the line is smaller due to the tax. Without a tax in place, the slope of the line is equal to the marginal propensity to consume (mpc). Once the tax is imposed, the slope of the line changes to (1-t) times the mpc as shown in Figure 18.12.

Figure 18.12: Introducing a Proportional Income Tax into the Consumption Function



This result is very different from the situation involving a lump sum tax. In that situation, explored in Chapter 13, the imposition of the lump sum tax causes a downward parallel shift of the consumption line because the vertical intercept falls.

In the case of a flat tax rate, the aggregate expenditure line also becomes flatter. We can derive the aggregate expenditures function (A) by adding together the after-tax consumption function (C_a) and the exogenously given values for investment spending (I), government spending (G), and net exports (X_n) as follows:

$$C_a = C_0 + mpc \cdot (1 - t)Y$$
$$I = I_0$$
$$G = G_0$$
$$X_n = X_{n0}$$
$$A = C_a + I + G + X_n = C_0 + mpc \cdot (1 - t)Y + I_0 + G_0 + X_{n0}$$
$$A = (C_0 + I_0 + G_0 + X_{n0}) + mpc \cdot (1 - t)Y$$

The aggregate expenditures function shows that the vertical intercept of the A curve is not affected when the flat tax is imposed. It simply becomes flatter just as the consumption function becomes flatter due to the smaller slope as shown in Figure 18.13.

Figure 18.13: Introducing a Proportional Income Tax into the Keynesian Cross Model

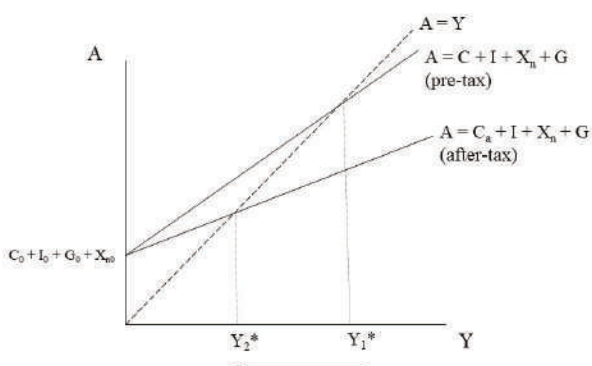


Figure 18.13 shows that the flatter A curve implies a lower equilibrium level of real output. The reader should

recall that the equilibrium output level in the Keynesian Cross model occurs where planned aggregate spending is equal to real output (A=Y). This condition is met where the aggregate expenditures curve intersects the 45-degree line in the graph. A higher tax rate would cause aggregate spending to fall even further and would further reduce real output and employment. On the other hand, a lower flat tax rate would represent an expansionary fiscal policy. It would make the A curve steeper and would raise the equilibrium output and level of employment.

The multiplier effect is also modified due to the imposition of a flat tax rate. Using the equilibrium condition that A = Y, the **government expenditures multiplier** may be derived in the following manner:

$$A = Y$$
$$(C_0 + I + G + X_n) + mpc \cdot (1-t)Y = Y$$
$$C_0 + I + G + X_n = Y - mpc \cdot (1-t)Y$$
$$C_0 + I + G + X_n = (1 - mpc \cdot (1-t))Y$$
$$Y = \frac{1}{1-mpc \cdot (1-t)}(C_0 + I + G + X_n)$$

In the equation for the equilibrium output, government spending, investment spending, and net export spending have been included as variables so that we may consider what happens when a change occurs in these components of spending. Specifically, we can derive the government expenditures multiplier as follows:

$$\Delta Y = \frac{1}{1-mpc \cdot (1-t)} \cdot \Delta G$$
$$\frac{\Delta Y}{\Delta G} = \frac{1}{1-mpc \cdot (1-t)}$$

Previously the government expenditures multiplier was equal to 1/(1-mpc). In the presence of a flat tax, the mpc is multiplied by (1-t). An increase in the tax rate will increase the denominator and lower the government expenditures multiplier. The result is intuitive. When additional government spending occurs, the additional income that households receive is only partly consumed because some is saved. The addition of a tax means that even less additional income is available for consumption, and the consequence is a lower multiplier.

The Implications of Different Taxation Systems for Macroeconomic Stability

The three systems of taxation that we have considered in this chapter carry very different implications for the stability of the economy. Table 18.6 summarizes the results that have been presented thus far regarding the different systems of taxation as well as claims about the degree of macroeconomic stability that each implies.[6]

Table 18.6: The Degree of Built-In Stability under Various Systems of Taxation

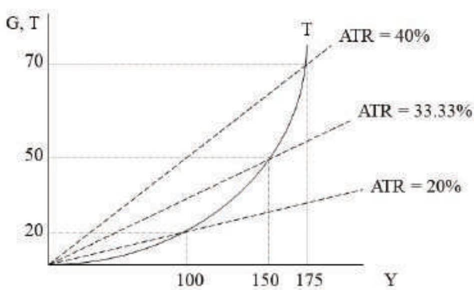
TAX SYSTEM	REAL GDP	AVERAGE TAX RATE	BUILT-IN STABILITY
Progressive	Increases (Decreases)	Increases (Decreases)	Most stable
Proportional	Increases (Decreases)	Constant (Constant)	Somewhat stable
Regressive	Increases (Decreases)	Decreases (Increases)	Least stable

Table 18.6 shows how the average tax rate changes as

real GDP (real income) changes under each of the three systems of taxation. It is asserted that progressive tax systems are the most stable taxation system and that proportional tax systems are somewhat stable. Regressive tax systems are regarded as the least stable system of taxation.

To understand how these conclusions are reached, we begin with the progressive tax system. Figure 18.14 shows how tax revenues change as real income increases.

Figure 18.14: Progressive Tax Systems (Rising Average Tax Rates)



In Figure 18.14, tax revenues rise very quickly as real income rises, which causes the ratio of taxes paid to income to grow quickly. The steeply rising T line implies a rising marginal tax rate (MTR) and a rising average tax rate (ATR). Each of these tax rates may be defined as follows:

$$MTR = \frac{\Delta T}{\Delta Y}$$

$$ATR = \frac{T}{Y}$$

The MTR in Figure 18.14 is reflected in the slope of the T line. It shows the additional taxes paid out of additional income. The ATR in Figure 18.14 is reflected in the slope of the ray from the origin drawn through the T line for a given income level. It should be clear that both the MTR and the ATR increase with income. The slope of the T line becomes steeper, which indicates a rising MTR. The rays drawn from the origin also become steeper, which implies a rising ATR.

If we add a government expenditures line to the graph of the steeply rising T line, then we obtain a graph like the one shown in Figure 18.15.

Figure 18.15: The Inherent Stability of a Progressive Tax System

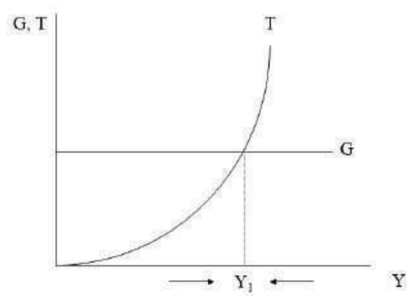


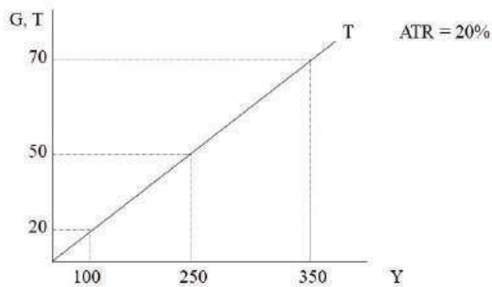
Figure 18.15 shows that the two curves intersect at Y_1, which implies that the budget is balanced with G equal to T. At income levels below Y_1, however, a government budget deficit exists. At income levels above Y_1 a government budget surplus exists. Now suppose that the

economy begins at income level, Y_1, and a recession occurs. Because budget deficits have an expansionary impact on the economy due to high government spending relative to tax revenues, the economy has an automatic tendency to expand. Similarly, if the economy begins at Y_1 and an expansion occurs, then the budget surplus that results has a contractionary impact on the economy due to low government spending relative to tax revenues.

We certainly should not view the Y_1 level of real output as an equilibrium level of real output. The equilibrium level of real GDP depends on many factors aside from the state of the government budget. Nevertheless, since the level of real output tends to increase in the case of a recession and tends to fall in the case of an expansion, the progressive tax system has a stabilizing effect on the economy. It should be noted that the budgetary response is automatic due to the automatic reduction in tax revenues when real income falls and the automatic increase in tax revenues when real income rises. The steep T line produces greater expansionary and contractionary effects. When real output falls, T falls significantly because the marginal tax rate declines. When real output rises, T rises significantly because the marginal tax rate increases. The resulting deficits and surpluses will, therefore, be larger as a result.

Let’s now turn to a proportional tax system or a flat tax system. Figure 18.16 shows how tax revenues change as real income increases.

Figure 18.16: Proportional Tax Systems (Constant Average Tax Rates)



In Figure 18.16, tax revenues rise at a constant rate relative to income, which causes the ratio of taxes paid to income to remain stable. The linear T line implies a constant marginal tax rate (MTR) and a constant average tax rate (ATR).

Again, the MTR in Figure 18.16 is reflected in the constant slope of the T line. The ATR in Figure 18.16 is reflected in the slope of the ray from the origin drawn through the T line for a given income level. In this case, a single ray from the origin passes through every point on the T line. Therefore, the slope of that ray represents the ATR at every level of income, which implies a constant ATR.

If we add a government expenditures line to the graph of the linear T line, then we obtain a graph like the one shown in Figure 18.17.

Figure 18.17: The Stabilizing Power of Proportional Tax Systems

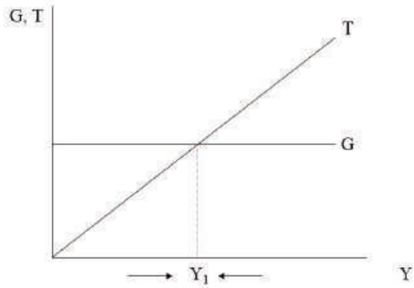
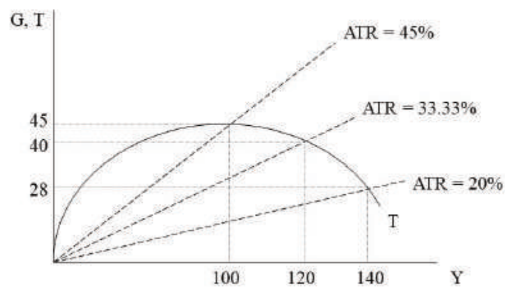


Figure 18.17 shows that the two curves intersect at Y_1, which implies that the budget is balanced with G equal to T. At income levels below Y_1, however, a government budget deficit exists. At income levels above Y_1 a government budget surplus exists. Now suppose that the economy begins at income level, Y_1, and a recession occurs. Because budget deficits have an expansionary impact on the economy due to high government spending relative to tax revenues, the economy has an automatic tendency to expand. Similarly, if the economy begins at Y_1 and an expansion occurs, then the budget surplus that results has a contractionary impact on the economy due to low government spending relative to tax revenues.

This result seems very much like the result we obtained for a progressive tax system. The difference is that tax revenues do not decline as rapidly during a recession, and they do not rise as quickly during an expansion. For this reason, we may consider proportional tax systems to have a stabilizing impact on the economy but less so than in the case of a progressive tax system.

Finally, let's consider a regressive tax system. Figure 18.18 shows how tax revenues change as real income increases.

Figure 18.18: Regressive Tax Systems (Falling Average Tax Rates)



In Figure 18.18, tax revenues rise and then fall as real income increases, which causes the ratio of taxes paid to income to fall. The inverted-U shape of the T line implies a falling MTR and a falling ATR.

Again, the MTR in Figure 18.18 is reflected in the falling slope of the T line. Notice that even when tax revenues are rising, the slope of the line is becoming flatter and so the MTR is falling.[7] That is, the additional taxes paid out of additional income are declining. The ATR in Figure 18.18 is reflected in the slope of the ray from the origin drawn through the T line for a given income level. In this case, the rays from the origin drawn through points on the T line are becoming flatter as real income rises. Therefore, the ATR falls as real income rises.

If we add a government expenditures line to the graph of the inverted-U T line, then we obtain a graph like the one shown in Figure 18.19.

Figure 18.19: The Destabilizing Power of Regressive Tax Systems

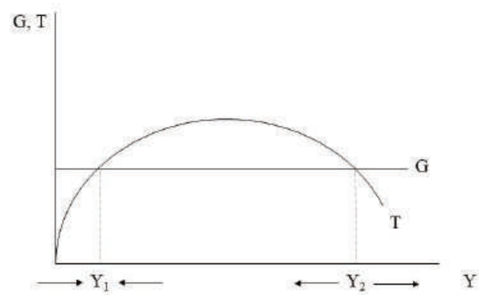


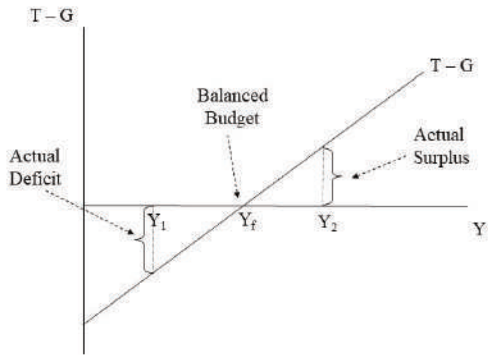
Figure 18.19 shows that the two curves intersect at two points, Y_1 and Y_2, which implies that the budget is balanced at both income levels with G equal to T. At income levels below Y_1, a budget deficit exists. At income levels above Y_1 a government budget surplus exists. These results are identical to what we have observed in the case of a progressive tax system and a proportional tax system. If the economy moves away from Y_1, then it has an automatic tendency to move back towards it.

Suppose, however, that the economy begins at income level, Y_2, and a recession occurs. In this case, a budget surplus results with tax revenues exceeding government spending, which is contractionary. The contractionary effect of the budget surplus is to move the economy further away from that output level and to worsen the recession. Similarly, if the economy begins at Y_2 and an expansion occurs, then a budget deficit will result with government spending exceeding tax revenues. Because budget deficits have an expansionary impact on the economy due to high government spending relative to tax revenues, the economy has an automatic tendency to expand, which might cause the economy to overheat. It should be clear that it is at least possible for a regressive tax system to destabilize the economy when the economy is operating at a level of output like Y_2. Critics sometimes point out that regressive tax systems tax lower income people at a higher rate than higher income people. We can add to that criticism that regressive tax systems also have greater potential to cause macroeconomic instability.

The Full Employment Budget

Another concept that neoclassical economists and Keynesian economists use to evaluate fiscal policy is the full employment budget. The **full employment budget**, as its name indicates, refers to the state of the budget at full employment. Neoclassical economists like to evaluate the nature of the government's fiscal policy at a point in time. If we only consider the **actual budget**, then the state of the economy might skew our perception. For example, consider the case of a balanced government budget at full employment as shown in Figure 18.20.[8]

Figure 18.20: A Balanced Full Employment Budget

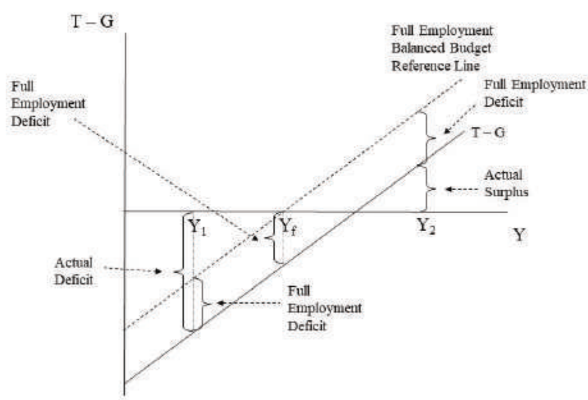


If the economy enters a recession and real output falls below the full employment output (Y_f), then an actual deficit will arise such as occurs at Y_1. Looking at the actual budget deficit suggests that the government is actively pursuing an expansionary fiscal policy. It is true that budget deficits have an expansionary effect on the economy, but this deficit exists only because of the automatic reduction in tax revenues that occurs when aggregate income declines. If we want to evaluate the government's fiscal policy separate from the impact of the **automatic stabilizers**, then it makes sense to look at the full employment budget, which suggests a neutral fiscal policy when the full employment budget is in balance.

Suppose that the economy experiences an expansion and real output rises to Y_2, which is above the full employment level of output. An actual surplus now exists, which might suggest a contractionary fiscal policy. It is true that budget surpluses have a contractionary effect on the economy, but this surplus exists only because of the automatic rise in tax revenues that occurs when aggregate income is rising. As in the case of an actual budget deficit, if we want to evaluate the government's fiscal policy separate from the impact of cyclical fluctuations, then we need to examine the full employment budget, which suggests a neutral fiscal policy due to the balanced full employment budget.

Now consider the case of a full employment budget deficit as shown in Figure 18.21.

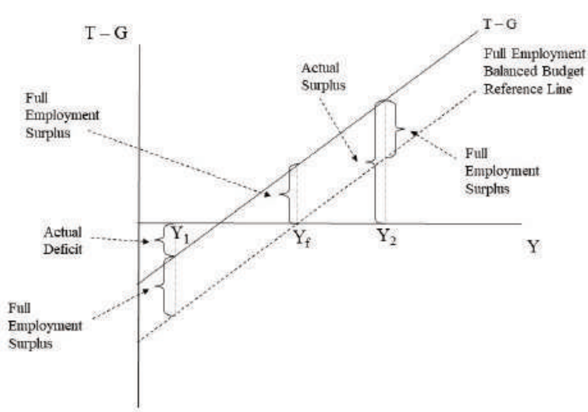
Figure 18.21: A Full Employment Budget Deficit



As Figure 18.21 shows, a budget deficit exists at the full employment level of output, Y_f. The line labeled T – G represents the actual budget whereas the **full employment balanced budget reference line** shows the state of the budget when it is balanced at full employment. The budget deficit at full employment suggests that the government is deliberately pursuing an expansionary policy. If the economy experiences a recession and real output falls to Y_1, however, then the actual budget deficit will grow as tax revenues fall. The full employment budget deficit is still represented as the difference between the actual budget line and the reference line, but now we can see an addition to the actual deficit. When a recession occurs, the actual deficit exceeds the full employment budget deficit. On the other hand, if the economy experiences an expansion and real output rises to Y_2, then an actual budget surplus will arise. Due to the full employment deficit, however, the actual surplus is smaller than it would be if the full employment budget was balanced.

Finally, consider the case of a full employment budget surplus as shown in Figure 18.22.

Figure 18.22: A Full Employment Budget Surplus



As Figure 18.22 shows, a budget surplus exists at the full employment level of output, Y_f. As before, the T-G line represents the actual budget whereas the full employment balanced budget reference line shows the state of the budget when it is balanced at full employment. The budget surplus at full employment suggests that the government is deliberately pursuing a contractionary policy. If the economy experiences an inflationary boom and real output rises to Y_2, however, then the actual budget surplus will grow as tax revenues rise. The full employment budget surplus is still represented as the difference between the actual budget line and the reference line, but now we can see an addition to the actual surplus. When an expansion occurs, the actual surplus exceeds the full employment budget surplus. On the other hand, if the economy experiences a recession and real output falls to Y_1, then an actual budget deficit will arise. Due to the full employment surplus, however, the actual deficit is smaller than it would be if the full employment budget was balanced. The concept of the full employment budget is useful because it allows us to see how an actual budget surplus or deficit may be larger or smaller than it would otherwise be due to the full employment budget gap.

Balanced Budget Amendments

Most U.S. states have **balanced budget amendments**, which is a constitutional requirement that the state governments balance their budgets each year. What are the macroeconomic consequences of adhering to such policies? Consider the case of a recession in which output falls below the full employment output as shown in Figure 18.23.

Figure 18.23: The Impact of a Balanced Budget Policy during a Recession

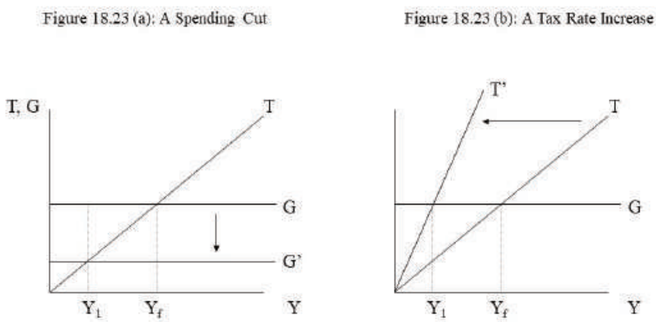


Figure 18.23 shows two methods of balancing the budget when output falls to Y_1 and a budget deficit arises. In Figure 18.23 (a), a government spending cut will restore balance to the budget since tax revenues have fallen. The problem with this approach is that a government spending cut is a contractionary fiscal policy, which is likely to worsen the recession. A balanced budget is thus achieved but at great cost to the nation's well-being. An alternative policy is to increase the average tax rate as shown in Figure 18.23 (b). In Figure 18.23 (b), tax revenues increase to restore balance to the budget, but a tax increase is also a contractionary policy taken during a recession. It seems that a balanced budget policy is likely to be a difficult and painful policy to pursue in a recession.

Now consider the case of an inflationary boom in which output rises above the full employment output as shown in Figure 18.24.

Figure 18.24: The Impact of a Balanced Budget Policy during a Period of Inflationary Expansion

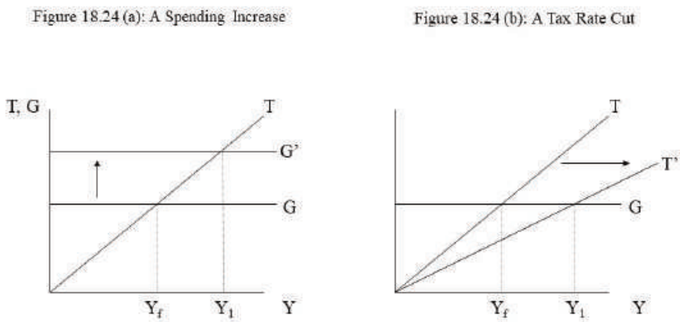


Figure 18.24 shows two methods of balancing the budget when output rises to Y_1 and a budget surplus arises. In Figure 18.24 (a), a government spending increase will restore balance to the budget since tax revenues have increased. The problem with this approach is that a government spending increase is an expansionary fiscal policy, which is likely to worsen the inflation. A balanced budget is once again achieved but at great cost to the nation's well-being. An alternative policy is to decrease the average tax rate as shown in Figure 18.24 (b). In Figure 18.24 (b), tax revenues fall to restore balance to the budget, but a tax decrease is also an expansionary policy taken during an inflationary boom. It seems that a balanced budget policy is likely to be a difficult and painful policy to pursue during an economic boom as well.[9]

A Comparison of Keynesian Full-Employment Policies and Austerity Policies

Much public discussion has been devoted to the proper role of government in the economy. Many argue for fiscal policy that promotes full employment while others advocate so-called **austerity measures** that reduce government budget deficits and government debt. Professor David Gleicher at Adelphi University has developed an excellent framework for demonstrating the difference between Keynesian full-employment policies and neoliberal austerity policies within a Keynesian model. The presentation here follows Prof. Gleicher's approach in its general outline.[10]

Let's first consider the case of an unemployment equilibrium outcome in the Keynesian Cross model. Assume that we know the following information about the economy:

$C_0 = 50$

$I_0 = 100$

$G_0 = 400$

$X_{n0} = 150$

$mpc = 0.75$

$t = 0.20$

Using this information, we write the aggregate expenditures function as follows:

$A = C_0 + mpc \cdot (1 - t)Y + I_0 + G_0 + X_{n0}$

$A = 50 + 0.75(1 - 0.20)Y + 100 + 400 + 150$

$A = 700 + 0.60Y$

Setting A equal to Y allows us to calculate the unemployment equilibrium level of real output (Y_u*) as follows:

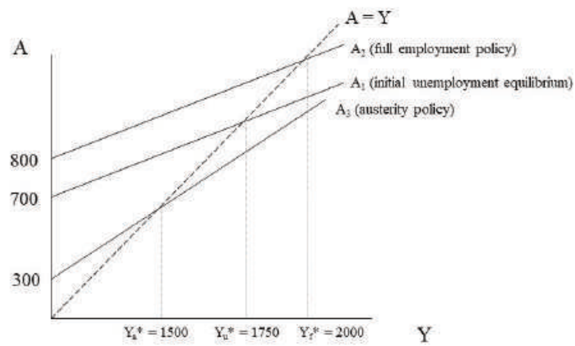
$700 + 0.6Y = Y$

$700 = 0.4Y$

$Y_u^* = 1750$

This solution is represented in Figure 18.25 as the initial unemployment equilibrium.

Figure 18.25: Full Employment Keynesian Policies versus Austerity Policies



We can also determine the state of the government budget at this unemployment equilibrium. Government spending is equal to 400, which is given information. Tax revenues may be calculated as follows:

$$T = tY = (0.20)(1750) = 350$$

Therefore, a budget deficit exists as shown below:

$$T - G = 350 - 400 = -50$$

When the economy is in a state of unemployment equilibrium, a debate erupts about what to do. Unemployment is high, and production has fallen. A large budget deficit exists as well. Political leaders and party officials begin to argue about the best path forward. Each side emphasizes different aspects of the nation's economic problems. One side points to job losses and economic insecurity while the other side points to government waste. One side advocates a full employment policy while the other side advocates austerity measures.

Consider the consequences of a Keynesian full employment policy. If the full employment level of real output (Y_f*) is estimated to be equal to 2000, then advocates of a full employment policy will favor increasing government spending so that the full employment level of real output becomes the new equilibrium for the economy. It is possible to use tax cuts or a combination of tax cuts and government spending to achieve this result, but for simplicity we focus only on the use of increased government spending. To determine the required level of government spending, we simply leave G as a variable and insert the full employment output (Y_f* = 2000) in the aggregate expenditures function as follows:

$A = 50 + 0.75(1 - .20)Y_F^* + 100 + G + 150$

$A = 50 + 0.75(1 - .20)(2000) + 100 + G + 150$

$A = 1500 + G$

Now we set aggregate spending equal to the full employment output:

$A = Y_F^*$

$1500 + G = Y_F^*$

$1500 + G = 2000$

$G = 500$

Because government spending must rise by 100 to make the full employment level of output the new equilibrium, the vertical intercept has increased to 800 as shown in Figure 18.25. We can also calculate the tax revenues as follows:

$T = tY = (0.20)(2000) = 400$

The government budget deficit in this case is equal to the following:
$T - G = 400 - 500 = -100$

The full employment policy has achieved full employment in this case, but it has also doubled the government budget deficit, which is why the advocates of austerity measures are so strongly opposed to this solution to the crisis.

Now let's consider the macroeconomic consequences of austerity measures. If austerity measures are implemented, then the government is committed to balancing the budget. To balance the budget, the following condition must hold:
$G = T = tY$

Government spending is now a function of real income. The reason is that when real income increases during an expansion, tax revenues rise and government spending must increase to maintain a balanced budget. Similarly, when real income falls during a recession, tax revenues fall and government spending must be cut to balance the budget.

If we substitute the new expression for government spending into the aggregate expenditures function, then we obtain the following:
$A = C_0 + mpc \cdot (1 - t)Y + I_0 + G_0 + X_{n0}$
$A = C_0 + mpc \cdot (1 - t)Y + I_0 + tY + X_{n0}$
$A = C_0 + (mpc \cdot (1 - t) + t)Y + I_0 + X_{n0}$
Because t has been added to the slope of the A line, it is steeper as shown in Figure 18.25. The intuition behind this addition to the slope is that when $1 of additional real income is received by households, tax revenues

increase by \$0.20 and the government must increase its spending by \$0.20 to maintain a balanced budget. Both government spending and consumer spending thus rise when real income rises, which creates a steeper aggregate expenditures line.

Let's now substitute the known information into the A function as follows:

$$A = 50 + (0.75(1 - 0.20) + 0.20)Y + 100 + 150$$

$$A = 300 + 0.80Y$$

Setting A = Y allows us to solve for the equilibrium outcome (Y_a*) as shown below:

$$300 + 0.80Y = Y$$

$$Y_a^* = 1500$$

The equilibrium output has fallen due to the austerity measures, which has worsened the recession, as shown in Figure 18.25. The government has managed to balance the budget, which can be confirmed as follows:

$$G = T = tY = (0.20)(1500) = 300$$

The budget is balanced and to achieve this result, the government reduced spending from 400 (the level at the initial unemployment equilibrium) to 300. Tax revenues also decline because of the reduction in real income. The balanced budget is achieved at great cost in terms of additional lost production and employment. This Keynesian analysis supports the view that no easy solution exists if one wants to return to full employment during a recession and balance the federal budget at the same time.

A Marxian Analysis of Government Borrowing and the Accumulation of Debt

In this final section, we will consider a Marxian analysis of the impact of government borrowing on the economy. Let's return to our earlier example involving the equalization of the profit rate across the industrial and financial sectors as shown in Table 18.7.

Table 18.7: A Numerical Example of Profit Rate Equalization across Sectors

Industrial Sector		Financial Sector	
Industrial Rate of Profit	25.00%	Financial Rate of Profit	25.00%
Non-Borrowed Capital ($)	305,000	Bank Operating Capital ($)	5,000
Borrowed Capital (Loan Capital) ($)	195,000	Bank Loan Capital ($)	35,000
Total Productive Capital ($)	500,000	Total Deposits ($)	200,000
Total Profit ($)	125,000	Reserve Requirement Ratio	20.00%
Total Interest Paid ($)	15,000	**Total Bank Capital ($)**	40,000
Total Annual Social Product ($)	625,000	**General Rate of Interest (Prime Rate)**	≈7.69%
		Loan Interest Received or Gross Bank Profit ($)	15,000
		Net Bank Profit ($)	10,000
	Aggregate Profits ($)	135,000	
	Total Social Capital ($)	540,000	
	General Rate of Profit (Gross)	25.00%	

In the example shown in Table 18.7, the market rate of interest has adjusted to equal the general rate of interest. The adjustment of the interest rate is what makes possible the equalization of the rate of profit in this example. Using this case as a starting point, we will consider two scenarios involving government borrowing and the consequences that it carries for the economy.

The first scenario involves government borrowing from the financial sector to finance the operation of state-owned enterprises (SOEs). The banks continue to lend $B_L+(1-R)D$, but now the loan is divided between the industrial sector and the state sector. Let's assume that a fraction, φ (between zero and one), is loaned to the industrial sector and that a fraction 1 – φ, is loaned to the state sector. Table 18.8 represents this case.

Table 18.8: The Government Using Borrowed Capital to Finance Production in State-Owned Enterprises (SOEs) with $\phi = 1/2$

State Sector		**Industrial Sector**		**Financial Sector**	
Industrial Rate of Profit	$p = 25\%$	Industrial Rate of Profit	$p = 25\%$	Financial Rate of Profit	$p = 25\%$
Borrowed Capital (Loan Capital) ($)	$K_{SB} = (1-\phi)(B_L+(1-R)D) = 97{,}500$	Non-Borrowed Capital ($)	$K_{INB} = 305{,}000$	Bank Operating Capital ($)	$B_O = 5000$
Total Profit ($)	$\pi_S = 24{,}375$	**Borrowed Capital (Loan Capital) ($)**	$K_{IB} = \phi(B_L+(1-R)D) = 97{,}500$	Bank Loan Capital ($)	$B_L = 35{,}000$
Total Interest Paid ($)	$i_g K_{SB} = 7{,}500$	**Total Productive Capital ($)**	$K_I = 402{,}500$	Total Deposits ($)	$D = 200{,}000$
Total State Product ($)	$TP_S = 121{,}875$	**Total Profit ($)**	$\pi_I = 100{,}625$	Reserve Requirement Ratio	$R = 20\%$
		Total Interest Paid ($)	$I_g K_{IB} = 7{,}500$	**Total Bank Capital ($)**	$B = 40{,}000$
		Total Private Product ($)	$TP_I = 503{,}125$	**General Rate of Interest (Prime Rate)**	$I_g = 7.69\%$
				Loan Interest Received or Gross Bank Profit ($)	$I_g K_{IB} + I_g K_{SB} = B_O + \Delta B = 15{,}000$
		Aggregate Profits ($)	$\pi_A = 135{,}000$	**Net Bank Profit ($)**	$\Delta B = 10{,}000$
		Total Social Capital ($)	$K_A = 540{,}000$		
		Total Social Product ($)	$TP_A = 625{,}000$		
		General Rate of Profit (Gross)	$p = 25\%$		

In Table 18.8 K_{SB} represents state-borrowed capital, π_S represents profits produced in the state sector, TP_S represents the total product of the state sector, K_{INB} represents non-borrowed capital in the industrial sector, K_{IB} represents industry-borrowed capital, K_I represents the total capital invested in industry, π_I represents the total profits of the industrial sector, and TP_I represents the total product of the industrial sector. When looking at the aggregate economy, K_A and π_A represent the aggregate capital and the aggregate profits across all three sectors.

Table 18.8 assumes that the fraction of the total loan capital borrowed in the industrial sector is ½ (i.e., φ = 1/

2). Similarly, the fraction of the total loan capital borrowed in the state sector is also ½ (i.e., $1 - \varphi = ½$). Because the non-borrowed capital in the industrial sector has not changed, the aggregate capital in the economy is the same. The total social product is also the same. The only important difference is that the composition of the total social product has changed with a portion of it now being produced in the state sector. Also, because the SOEs in the state sector hire workers to produce commodities, this sector creates surplus value and appropriates profits. This situation thus represents a Marxian example of complete crowding out of private sector activity. Even so, the situation has not changed for the working class. Workers still create profits. The only difference is that the exploiters have changed to include both industrialists and state managers. Finally, a portion of the profits of the SOEs is paid as interest to the banks that granted the loans to the government in the first place. The government can also repay the entire debt (the principal amount of the loan) because it has received revenue from the sale of commodities equal to the value of the total product in that sector.

The second scenario that we will consider involves government borrowing from the financial sector. This time, however, the government does not use the borrowed funds as capital. Instead, it plans to spend the entire amount on commodities produced in the industrial sector. This borrowing withdraws a great deal of capital from the economy. If nothing else changes, the result will be a major contraction of economic activity and inflation as the prices of the remaining commodities are driven up. In fact, this scenario can lead to many different results depending on the assumptions that we make about the private sector's response to the government borrowing, and we will only consider one special case.

Let's assume that the government enters into private sector contracts immediately with plans to use its borrowed funds to purchase commodities from the industrial sector. Let's further assume that the government contracts encourage the introduction of more non-borrowed capital into the industrial sector as those with hoards decide that the government guarantees justify new investment in this sector. The amount of capital that will be introduced into the industrial sector in response to the government contracts (K_{GC}) may be calculated using the following equation:

$$(1-\phi)\{B_L+(1-R)D\}=(1+p)K_{GC}$$

The left-hand side in this equation represents the funds that the government has borrowed from the financial sector, which it spends on commodities produced in the industrial sector. The right-hand side represents the value of output produced in the industrial sector under government contract. If the fraction of the loan capital that the government borrows is ½ (i.e., $1 - \varphi = 1/2$), then this case may be represented as in Table 18.9.

Table 18.9: The Government Using Borrowed Capital to Purchase Privately Produced Commodities from Industry with $\phi = 1/2$

Industrial Sector (for government consumption)		Industrial Sector (for private consumption)		Financial Sector	
Industrial Rate of Profit	$p = 25\%$	Industrial Rate of Profit	$p = 25\%$	Financial Rate of Profit	$p = 25\%$
State Non-Borrowed Capital ($)	$K_{GC} = 78{,}000$	Non-Borrowed Capital ($)	$K_{NB} = 305{,}000$	Bank Operating Capital ($)	$B_O = 5000$
Total Profit ($)	$pK_{GC} = 19{,}500$	**Borrowed Capital (Loan Capital) ($)**	$K_{IB} = \phi(B_L + (1-R)D) = 97{,}500$	Bank Loan Capital ($)	$B_L = 35{,}000$
Total Interest Paid ($)	$i_g(K_{GC} + \pi_{GC}) = 7{,}500$	**Total Productive Capital ($)**	$K_{PC} = 402{,}500$	Total Deposits ($)	$D = 200{,}000$
Product for the State ($)	$TP_{GC} = 97{,}500$	**Total Profit ($)**	$\pi_{PC} = 100{,}625$	Reserve Requirement Ratio	$R = 20\%$
		Total Interest Paid ($)	$i_g K_{IB} = 7{,}500$	**Total Bank Capital ($)**	$B = 40{,}000$
		Total Private Product ($)	$TP_{PC} = 503{,}125$	**General Rate of Interest (Prime Rate)**	$i_g = 7.69\%$
				Interest from Loans to the State ($)	$i_g(K_{GC} + \pi_{GC}) = 7{,}500$
Aggregate Wages ($)	$W_A = 347{,}000$	**Aggregate Profits ($)**	$\pi_A = 130{,}125$	**Interest from Private Sector Loans ($)**	$i_g K_{IB} = 7{,}500$
		Total Social Capital ($)	$K_A = 520{,}500$	**Total Loan Interest Received ($)**	$B_O + \Delta B = 15{,}000$
		Total Social Product ($)	$TP_A = 600{,}625$	**Net Bank Profit ($)**	$\Delta B = 10{,}000$
		General Rate of Profit (Gross)	$p = 25\%$		

We may calculate the non-borrowed capital in the industrial sector (K_{GC}) in Table 18.9 as follows:

$$(1 - \tfrac{1}{2})\{35{,}000 + (1 - 0.20)200{,}000\} = (1 + 0.25)K_{GC}$$
$$K_{GC} = 78{,}000$$

The total profit in the government-contracted industrial sector (π_{GC}) may be calculated as follows:

$$\pi_{GC} = pK_{GC} = (0.25)(78{,}000) = 19{,}500$$

Table 18.9 also includes the total product in the government-contracted industrial sector (TP_{GC}), the profits in the private sector-contracted industrial sector (π_{PC}), and the total product in the private sector-contracted industrial sector (TP_{PC}). It should be clear from a comparison of Table 18.8 and Table 18.9 that the total social product has fallen. The reason is that the government spends the entire borrowed funds rather than advances it as capital. Therefore, the new capital advanced in industry due to the government contracts is only a fraction of the government's borrowed amount. The industrialists who enter into government contracts expect to receive the average profit in return for their investments and so they only advance as much capital as will enable them to recover the investment plus the average profit. The government borrowing thus leads to a minor contraction in the private sector. The government can borrow from many sources, however, including from sources that do not reduce the capital invested in the private sector, and so the government demand for commodities will generally increase the total social product. We have assumed that the government borrowing partially interferes with private sector productive activity to simplify the example and the calculations in this section.

Finally, we need to consider the matter of the interest that the government owes to the financial sector for its loan (I_G). It uses the entire borrowed amount of $97,500 to purchase the total product of the government-contracted industrial sector. The government owes interest equal to the following:

$$I_G = i_g(K_{GC} + \pi_{GC}) = (0.0769)(97,500) = 7,500$$

The government cannot pay the interest out of profits as it did in the first scenario when it advanced the funds as capital and appropriated the profits of the SOEs. To pay the interest, it must impose a tax on profits, on wages, or on a combination of wages and profits. Let's suppose that the government taxes all profits at the same rate so that it can exactly meet its interest payment. In that case, the tax rate on profits (t_π) is calculated using the aggregate profits across all three sectors (π_A).

$t_\pi \pi_A = I_G$

$t_\pi = \frac{I_G}{\pi_A}$

$t_\pi = \frac{7{,}500}{130{,}125} = 5.76\%$

This tax rate on profits will ensure that the government receives just enough tax revenue to equal the interest owed on its debt.

Suppose the government taxes only wages to acquire enough tax revenue to pay the interest on its debt. To consider this case, we will assume that 2/3 of the aggregate capital stock (K_A) consists of wages. Then the aggregate wages across all three sectors (W_A) may be calculated as follows:

$W_A = (\frac{2}{3})(K_A) = 347,000$

The tax rate on wages (t_W) may then be calculated as follows:

$t_w W_A = I_G$

$t_w = \frac{I_G}{W_A}$

$t_w = \frac{7{,}500}{347{,}000} = 2.16\%$

This tax rate on wages will ensure that the government receives just enough tax revenue to equal the interest owed on its debt. It is worth noting that the tax on wages reduces the after-tax wages that workers receive. Workers thus have less money to purchase the commodities they need to reproduce their labor-power each day. Because the value of labor-power is not a biological minimum, it is possible that after-tax wages may fall below the value of labor-power. Alternatively, the lower after-tax wages may represent a change in the value of labor-power such that workers are no longer recognized as requiring the previously affordable larger bundle of commodities for their daily consumption. Either way, workers experience a reduction in their living standards due to the tax.

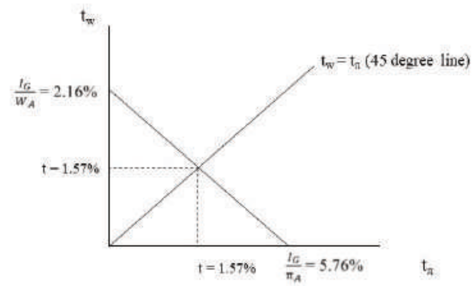
A final possibility is that the government imposes a tax on both wages and profits. In this case, we simply add the two sources of tax revenue and set the total tax revenue equal to the interest owed.
$t_w W_A + t_\pi \pi_A = I_G$

If all wages and profits in the economy are taxed at the same rate (t), then we calculate the following:
$tW_A + t\pi_A = I_G$
$t(W_A + \pi_A) = I_G$
$t = \frac{I_G}{W_A + \pi_A}$
$t = \frac{7{,}500}{347{,}000 + 130{,}125} = 1.57\%$
Taxing all wages and profits the same allows all income to be taxed at a lower rate than if only wages or profits are taxed. Of course, because workers are responsible for creating all the new value (i.e., wages plus profits), if their wages are taxed at all, then they not only have the profits they produced taken from them, but some of their wage income as well.

It is possible to see all the combinations of tax rates on profits and wages if we consider the general equation that allows for a combination of profit taxes and wages taxes and then solve for the tax rate on wages as follows:
$t_w W_A + t_\pi \pi_A = I_G$
$t_w = \frac{I_G}{W_A} - \frac{\pi_A}{W_A} t_\pi$

Figure 18.26 shows all the combinations of the two tax rates that will generate just enough tax revenue to ensure that the government can meet its interest payment on the debt.

Figure 18.26: Combinations of Tax Rates that Allow the Government to Pay the Interest Owed on its Debt



As Figure 18.26 shows, if only wages are taxed, then the tax rate on wages may be calculated simply by dividing the interest owed by the aggregate wage income. If only profits are taxed, then the tax rate on profits may be calculated simply by dividing the interest owed by the aggregate profit income. If both wages and profits are taxed at the same rate (t), then the tax rate is calculated by dividing the interest owed by the sum of all wages and profits. Finally, the slope indicates that a one percent increase in the tax rate on profits will cause a reduction in the tax rate on wages of π_A/W_A.

We can conclude this section with a look at the state of the government budget and the government debt in this second scenario. Government outlays (G) include the state spending on commodities (TP_{GC}) plus the interest paid on the loan (I_G).

$$G = TP_{GC} + I_G$$

Government receipts (R) only include tax revenues.

$$R = t_w W_A + t_\pi \pi_A$$

The government budget gap is equal to the following (since taxes equal interest owed):

$$R - G = t_w W_A + t_\pi \pi_A - (TP_{GC} + I_G) = -TP_{GC} = -(1-\phi)\{B_L + (1-R)D\}$$

In other words, a budget deficit exists that is equal to the borrowings from the financial sector. This amount also represents the increase in the government debt for this period.

Following the Economic News[11]

On December 22, 2017, President Donald Trump signed the Tax Cuts and Jobs Act into law. The law provides tax cuts for corporations, small businesses, and individual taxpayers. Among other changes to the federal tax code, it includes a cut in the corporate income tax rate from 35% to 21% and a cut in the marginal income tax rates for five of the seven income tax brackets. *The Wall Street Journal* has reported that the immediate impact of the new law in early 2018 was to reduce tax withholding from paychecks and make possible one-time bonuses for some households as businesses responded to the tax cuts. Large tax cuts represent an expansionary fiscal policy measure and one would expect a rise in aggregate demand due to the tax changes. Nevertheless, the WSJ reported in early March that the income increases have not yet resulted in higher consumer spending, which might slow the growth of aggregate output in the first quarter of 2018. The WSJ asserts that a lag is responsible and that it may take a few months for consumer spending to rise in response to the tax cuts. The inflation rate was also reported as lower than the 2% annual inflation target of the Fed. At the same time, the core inflation rate was relatively high, and the WSJ expects the Fed to be "slightly more aggressive" in raising interest rates in 2018. This recent analysis suggests that expansionary fiscal policy can experience a delay in terms of its impact on consumer spending and aggregate output. With the economy close to the full employment level and the recent uptick in the price level, however, the Fed is likely to pursue an offsetting contractionary monetary policy to prevent demand-pull inflation later in the year. Using our Marxian perspective, it is also obvious that the tax cuts on wage income and profit income are sure to increase the federal deficit and the federal debt. Indeed, the Congressional Budget Office (CBO) estimates that the federal debt is expected to rise by close to $1.5 trillion over the next 10 years due to the tax cut.

Summary of Key Points

1. A budget deficit exists when government outlays exceed government receipts for the year; a budget surplus exists when government receipts exceed government outlays for the year; and a balanced budget exists when government outlays equal government receipts for the year.

2. The unified budget deficit (or surplus) is equal to the sum of the on-budget deficit (or surplus) and the off-budget deficit (or surplus).

3. Federal outlays consist of appropriated programs, mandatory spending, and net interest.

4. Federal receipts consist of individual income taxes, corporate income taxes, payroll taxes, unemployment insurance taxes, excise taxes, estate and gift taxes, customs duties, and profit distributions from the Federal Reserve.

5. The federal debt represents the accumulated debt of the federal government minus the amount that has been repaid over the years.

6. Expansionary fiscal policy uses tax cuts and/or government spending increases to combat recession. Contractionary fiscal policy uses tax increases and/or government spending cuts to combat inflation.

7. When the government borrows to finance deficit spending, the consequence can be complete crowding out, partial crowding out, or no crowding out of private investment and private consumption. Net exports may also be negatively affected depending on the exchange rate policy of the central bank.

8. When the government runs a budget surplus and repays its debt, the consequence can be demand-pull inflation.

9. Marginal tax rates refer to the rate at which an additional dollar of income is taxed. Average tax rates refer to the ratio of total taxes paid to total income.

10. In progressive tax systems, the average tax rate rises as income rises; in a proportional tax system, the average tax rate remains constant as income changes, and in a

regressive tax system, the average tax rate falls as income rises.

11. When a flat tax rate is introduced into the Keynesian Cross model, the government expenditures multiplier changes to 1/[1-(1-t)mpc], and the equilibrium output falls.

12. A progressive tax system is the most stable tax system; a proportional tax system is the second most stable tax system; and a regressive tax system is the least stable tax system.

13. The full employment budget allows us to consider the state of the government budget at full employment and ignore the impact on the budget of automatic changes in tax revenues that occur over the course of the business cycle.

14. Strict adherence to a balanced budget policy tends to worsen recessions and inflationary booms.

15. Full employment Keynesian policies during a recession can generate full employment but cause large budget deficits. Austerity measures during a recession can balance the budget but cause a further contraction of aggregate output and employment.

16. In a Marxian framework, when the government borrows to finance production in SOEs, the only macroeconomic consequence is a change in the composition of the total output. When the government borrows to purchase privately produced output, however, it accumulates debt and must impose taxes on wages or profits (or both) to pay interest on its debt.

List of Key Terms

Fiscal policy

Balanced budget

Budget deficit

Budget surplus

Unified budget deficit

Unified budget surplus

On-budget deficit

On-budget surplus

Off-budget deficit

Off-budget surplus

Appropriated programs (discretionary programs)

Mandatory spending

Entitlement programs

Net interest

Federal debt

Expansionary fiscal policy

Contractionary fiscal policy

Complete crowding out

Partial crowding out

Complete accommodation

Flexible exchange rates

Fixed exchange rates

Marginal tax rates

Income tax brackets

Median income

Average tax rate

Progressive tax system

Proportional tax system

Flat tax

Regressive tax system

Government expenditures multiplier

Full employment budget

Actual budget

Automatic stabilizers

Full employment balanced budget reference line

Balanced budget amendments

Austerity measures

Problems for Review

1. Suppose that individual income tax revenues are $1.2

trillion, corporate income taxes are \$0.5 trillion, and excise taxes are \$0.2 trillion. Social Security payroll taxes (off-budget) are \$0.75 trillion. National defense spending is \$0.6 trillion, spending on Human Resources is equal to \$1.75 trillion, and net interest equals \$0.4 trillion. Finally, assume that off-budget Social Security expenditures are \$0.8 trillion and spending by the Postal Service amounts to \$0.1 trillion. Calculate the following:

a. On-budget receipts
b. On-budget outlays
c. Off-budget receipts
d. Off-budget outlays
e. On-budget deficit or surplus
f. Off-budget deficit or surplus
g. Unified budget deficit or surplus

2. Suppose that the marginal propensity to consume (mpc) is 0.8. Also assume that government spending increases by \$200 billion and lump sum taxes fall by \$100 billion. What is the total change in the equilibrium real GDP, if the price level is fixed in the short run?

3. Suppose you are a single taxpayer and that your annual income is \$175,000. Calculate the total taxes that you owe the federal government for the year using Table 18.5. Also calculate your average tax rate (ATR). How does it compare with the marginal tax rate in your income tax bracket? What is the reason for this relationship?

4. Suppose you know the following information about the economy:

a. Autonomous consumer spending is \$200 billion.
b. Autonomous investment spending is \$100 billion.
c. Autonomous government spending is \$150 billion.

d. Autonomous net export spending is \$75 billion.
e. The flat tax rate is 15% or 0.15.
f. The marginal propensity to consume is 0.7.

Solve for the equilibrium output. Then place the aggregate expenditures function on a graph with A on the vertical axis and real GDP (Y) on the horizontal axis. Include the 45-degree line (A = Y) on your graph. Identify the numerical values for the equilibrium output and the vertical intercept on the graph.

5. Suppose real GDP rises from \$2 trillion to \$3.5 trillion, and tax revenues rise from \$0.25 trillion to \$0.6 trillion. Calculate the average tax rate before and after the change. What happens to it? What kind of tax system is it?

6. Is it possible for a government to have a full employment budget surplus but then also to be running an actual deficit? What would need to happen to bring about that situation? Consider Figure 18.22 when answering this question.

7. Use the information from question 4 for these problems.

a. Suppose that the full employment output is \$1500 billion. What level of government spending will generate this level of output as the equilibrium output? What will tax revenues be at this output level? What will be the state of the government budget at this output level?

b. Suppose the government implements austerity measures in the hopes of balancing the budget. What should it do? Find the new equilibrium output. Does the result surprise you? How is this outcome possible?

8. Consider Figure 18.26. Answer the following questions:

a. What kind of change might cause a parallel shift of the tax rate line to the right?

b. How would a change in the distribution of income between profits and wages affect the tax rate line? Consider, for example, a rise in profits and a fall in wages. Assume that the sum of wages and profits remains the same. Redraw the graph in Figure 18.26, and then add the new tax rate line on the same graph.

Notes

1. Office of Management and Budget. Tables 2.1, 2.5, 3.1, and 3.2. Historical Tables. Web. Accessed on February 21, 2018. https://www.whitehouse.gov/omb/historical-tables/
2. Office of Management and Budget. Tables 2.1, 2.5, 3.1, and 3.2. Historical Tables. Web. Accessed on February 21, 2018. https://www.whitehouse.gov/omb/historical-tables/
3. See Bade and Parkin (2013), p. 827, for an explanation of how changes in government spending, transfer payments, and taxes may be used separately or in combination with each other as part of a nation's fiscal policy. McConnell and Brue (2008), p. 209-211, also discuss the effects of changes in government spending and taxes, separately and then in combination with one another, where the policies reinforce one another in their impact on real output. Samuelson and Nordhaus (2001), p. 504, discuss the combined impact of changes in taxes and government spending also, but they consider the case where the two policies offset each other in their impact on real output. Chisholm and McCarty (1981), p. 175-178, discuss expansionary and contractionary spending policies and then

expansionary and contractionary tax policies before considering the synthesis case.

4. Wolff and Resnick (2012), p. 114-115, emphasize this point.
5. As Chisholm and McCarty (1981), p. 187-188, explain, debt retirement may be "fiscally counterproductive" since the aim of the budget surplus is to "diminish total demand."
6. McConnell and Brue (2008), p. 212-213, evaluate the three systems of taxation based on their degrees of built-in stability. This section greatly expands upon their discussion.
7. Eventually, the slope of the T line becomes negative, which implies a negative marginal tax rate. This case might seem unrealistic as it seems to imply negative marginal income taxes for the rich (i.e., subsidies) as their income grows. The MTR here, however, only shows what happens to total taxes as income grows. If enough taxpayers move to higher income levels where the marginal tax rates are lower (albeit still positive), then the aggregate MTR falls as shown.
8. Graphs that show how the budget gap changes relative to changes in aggregate output may be found in Solomon (1964), p. 107, and in Oakland (1969), p. 350.
9. Neoclassical textbooks tend to emphasize the difficulties associated with balanced budget amendments. For example, see Hubbard and O'Brien (2019), p. 970-971, and Chiang and Stone (2014), p. 564.
10. I am deeply grateful to Prof. Gleicher for granting me permission to include a summary of his model in this section, which uses a different numerical example than the one he originally used to illustrate these points. The original source is: Gleicher, David. "A Novel Method of Teaching Keynesian Demand versus Neoliberal Austerity Policies within the Simple Keynesian Model." Union for Radical Political Economics (URPE) Newsletter. Volume 44, Number 3, Spring 2013: 6-7. The final, definitive version is available at https://urpe.org/

content/media/UA_URPE_Past_Newsletters/
spring2013newsletter.pdf.

11. Morath, Eric. "U.S. After-Tax Incomes Rise Due to Tax-Code Changes, Spending Slows." *The Wall Street Journal*. Web. Updated March 1, 2018. Accessed on March 6, 2018.

PART V

PART FOUR: PRINCIPLES OF INTERNATIONAL ECONOMIC THEORY

CHAPTER 19

THEORIES OF INTERNATIONAL TRADE

Goals and Objectives:

In this chapter, we will do the following:

1. *Explain* the neoclassical theory of comparative advantage
2. *Analyze* the limits to the terms of trade and the determination of the terms of trade
3. *Explore* criticisms of the theory of comparative advantage
4. *Investigate* the so-called New Trade Theory
5. *Examine* the neoclassical approach to tariffs and quotas
6. *Consider* radical theories of world trade

In this final part of the book, we explore key principles that will help us understand the operation of the world economy. This chapter concentrates on theories of international trade. The final chapter of the book considers theories of international finance. As in other chapters, much of the focus in the present chapter will be on the dominant neoclassical perspective. Unlike other economics principles textbooks, however, we will also consider criticisms of the mainstream theory of trade as well as radical theories that serve as alternatives to neoclassical trade theory. As before, these alternative theories of trade will provide us with a different perspective of the same material, but they will also deepen our understanding of mainstream trade theory.

The Ricardian Theory of Comparative Advantage

The theory of international trade that dominates mainstream neoclassical discourse today has its roots in a theory developed by David Ricardo in his 1817 book *The Principles of Political Economy and Taxation*. According to Ricardo, two nations can always gain from trade as long as each nation has a **comparative advantage** in the production of at least one commodity. That is, he argued that a richer nation might still gain from trade with a poorer nation even if it is a better producer of every commodity that the two countries produce. If the rich nation is relatively much better at producing one commodity than another, then it makes sense for the rich nation to specialize in the production of the commodity that it is relatively much better at producing. It can then engage in trade with the poor country, and both nations will be better off.

When Ricardo presented his theory, he did so using the classical labor theory of value. He assumed that one nation was absolutely better at producing both commodities. Within that framework, a nation has an **absolute advantage** if it can produce one unit of a commodity with less labor time than another nation. A nation has a comparative advantage in the production of a commodity, on the other hand, if its absolute advantage is relatively greater for that commodity. In that case, it should specialize in the commodity in which it has a comparative advantage. The other country should specialize in the production of the other commodity in which it will have a comparative advantage. Even though the poor country has an absolute *disadvantage* in the production of both commodities (more labor time is required to produce a unit of each commodity), it nevertheless has a smaller absolute disadvantage in the production of one commodity. The commodity in which it has a smaller absolute disadvantage is the one in which it has a comparative advantage.

Table 19.1 provides an example (like Ricardo's original example) of two nations and the amounts of labor required to produce one unit of cloth and one unit of wine.

Table 19.1: A Ricardian Presentation of Comparative Advantage Theory

	Wine (labor hours to produce 1 unit)	Cloth (labor hours to produce 1 unit)
Portugal	5	6
England	3	2

Table 19.1 shows that England has an absolute advantage in the production of both wine and cloth. That is, it requires absolutely less labor time in England to produce each commodity than in Portugal. At the same time, it requires 1 and 2/3 times as much labor time to produce wine in Portugal (= 5/3) as it does in England. Similarly,

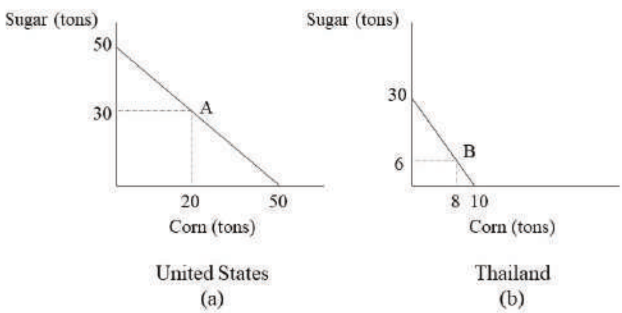
it requires 3 times the amount of labor time in Portugal to produce cloth (= 6/2). Therefore, Portugal's absolute disadvantage is greatest in cloth production and so we should expect it to have a comparative advantage in wine production.

Similarly, England requires only 0.6 times as much labor time as Portugal to produce wine (= 3/5). England also requires only about 0.33 times as much labor time as Portugal to produce cloth (= 2/6). Again, we see that England can produce each commodity in less time than Portugal. Nevertheless, England's absolute advantage is clearly greatest in the production of cloth. Therefore, England's comparative advantage is in the production of cloth. In conclusion, England should specialize in cloth production, Portugal should specialize in wine production, and the two nations can trade to the benefit of each.

The Modern Theory of Comparative Advantage: The Case of Constant Opportunity Cost

Of course, mainstream economists have long since abandoned the classical labor theory of value. As a result, the Ricardian presentation is not the best way to represent the mainstream theory of international trade. Instead, we will return to the production possibilities model and consider an example in which Thailand and the United States produce both sugar (S) and corn (C). Each nation has a given production technology and given stocks of land, labor, and capital. We will also modify our earlier assumption from Chapter 2 that each nation experiences **increasing marginal opportunity costs**. The reader might recall that increasing opportunity cost arises because societies generally possess **heterogeneous resources**. Instead, we will assume that each nation possesses **homogeneous resources** that are equally well-suited to all lines of production and thus are easily transferrable from one industry to another. The assumption of homogeneous resources generates a pattern of constant marginal opportunity costs. Figure 19.1 shows the production possibilities for the U.S. and Thailand.

Figure 19.1: Production Possibilities Frontiers for Two Nations



Because the slope of the production possibilities frontier (PPF) signifies marginal opportunity cost, the assumption of constant marginal opportunity costs explains the linear production possibilities frontiers of each nation. Points A and B on the two PPFs are included simply to provide examples of points that lie on the curves and to remind the reader that a nation may choose to be at any specific point on the PPF at a point in time by shifting resources across industries.

Given this different way of representing the theory of comparative advantage, we need to modify our earlier definitions to fit this framework. That is, a nation will be said to have an *absolute advantage* if it can produce absolutely more of a specific commodity during a given time period. For example, if the U.S. shifts all its resources to sugar production, then it can produce 50 tons of sugar. In that case, the economy would operate at the vertical intercept in Figure 19.1 (a). At the same time, if Thailand shifted all its resources to sugar production, then it could produce 30 tons of sugar. The economy would also be operating at the vertical intercept in Figure 19.1 (b). Because the U.S. can produce absolutely more sugar than Thailand, the U.S. has an absolute advantage in sugar production. Similarly, if all resources were devoted to corn production, then both countries would operate at the horizontal intercepts in Figure 19.1. The U.S. would produce 50 tons of corn, and Thailand would produce 10 tons of corn. Once again, the U.S. has an absolute advantage in corn production.

Because the U.S. has an absolute advantage in the production of both corn and sugar, the conclusion might be drawn that the U.S. should not trade with Thailand. If we recall Ricardo's insight, however, that a nation only needs to have a comparative advantage for trade to be beneficial to it, then trade might still be worthwhile. To provide a definite answer to this question, we must introduce a modified definition of comparative advantage that is compatible with the production possibilities framework. It will be said that a nation possesses a *comparative advantage* in the production of a commodity if the marginal opportunity cost of production is lower for that nation than for another nation.

As mentioned previously, the marginal opportunity cost, also known as the **domestic terms of trade**, is reflected in the slope of the PPF. In the U.S., therefore, the marginal opportunity cost of corn is the following:

$$\frac{\Delta S}{\Delta C} = -\frac{50S}{50C} = -\frac{1S}{1C}$$

In other words, the opportunity cost of producing 1 unit of corn is 1 unit of sugar in the U.S. In Thailand, however, the domestic terms of trade are rather different. The marginal opportunity cost of producing corn in Thailand is the following:

$$\frac{\Delta S}{\Delta C} = -\frac{30S}{10C} = -\frac{3S}{1C}$$

That is, the marginal opportunity cost of producing 1 unit of corn in Thailand is 3 units of sugar. It should be clear that the marginal opportunity cost of producing corn is lower in the United States. In the U.S. it only costs 1 unit of sugar rather than 3 units as in Thailand. Therefore, the U.S. has a comparative advantage in the production of corn.

Thailand must have a comparative advantage in sugar production, but to verify this result, let's consider the reciprocal of the slope. The reciprocal of the slope will allow us to see the marginal opportunity cost of producing 1 unit of sugar. For example, in the U.S. the marginal opportunity cost of sugar production is the following:

$$\frac{\Delta C}{\Delta S} = -\frac{50C}{50S} = -\frac{1C}{1S}$$

In other words, the marginal opportunity cost of 1 unit of sugar is one unit of corn. Similarly, we can write the marginal opportunity cost of 1 unit of sugar in Thailand as follows:

$$\frac{\Delta C}{\Delta S} = -\frac{10C}{30S} = -\frac{1/3C}{1S}$$

In other words, the marginal opportunity cost of 1 unit of sugar in Thailand is 1/3 units of corn. Therefore, the marginal opportunity cost of sugar production is lower in Thailand than in the United States since it costs 1 unit of corn in the U.S. but only 1/3 units of corn in Thailand to produce 1 unit of sugar. Thailand thus has a comparative advantage in sugar production.

In such examples, it will always be the case that if one nation has a comparative advantage in the production of a commodity, then the other nation will have a comparative advantage in the production of the other commodity. The only case in which a nation will not have a comparative advantage in the production of either commodity is the one in which the marginal opportunity cost is the same in the two nations. Trade in that case will not achieve anything that cannot be achieved domestically.

In this case, however, trade can benefit both nations. To understand how, we assume that each nation completely specializes in the production of the commodity in which it has a comparative advantage. That is, it is assumed that the U.S. shifts all its resources to the production of corn since it has a comparative advantage in corn production. Similarly, Thailand shifts all its resources to sugar production since that is where its comparative advantage lies. In Figure 19.1, the U.S. will produce at the horizontal intercept in (a), and Thailand will produce at the vertical intercept in (b).

Now that each nation is producing only one commodity, it will certainly want to trade a part of its production for the other commodity that is now only produced by its trading partner. The amount of the other commodity that each country will wish to import and the amount of its own commodity that it wishes to export will depend on the preferences of the consumers in each nation. Because the PPF only tells us about production possibilities and does not tell us about consumption possibilities, we do not possess enough information to address this question.

We are in a position, however, to draw some conclusions regarding the **international terms of trade** that will be established in the world market. The international terms of trade tell us how much sugar Thailand is willing to exchange for 1 unit of American corn. Its reciprocal tells us how much corn the U.S. is willing to exchange for 1 unit of Thai sugar. This international rate of exchange will be determined through a process of negotiation between the two nations. That is, competition in the marketplace will ultimately determine the price that emerges for each commodity.

Certain **limits to the terms of trade** exist though. To understand why such limits exist, consider the situation from Thailand's perspective first. Would Thailand ever trade more than 3 units of sugar for 1 unit of corn? The answer is definitely "no." According to its domestic terms of trade, Thailand can produce 1 unit less of corn and 3 units more of sugar by shifting resources from corn production to sugar production. It should never pay more in the world market than 3 units of sugar for 1 unit of corn since it is capable of that tradeoff at home. Similarly, the U.S. will never pay more in the world market than 1 unit of corn for 1 unit of sugar. According to its domestic terms of trade, the U.S. can shift resources from corn production to sugar production. If it does so, then it will gain 1 unit of sugar at the cost of 1 unit of corn. Since it can obtain 1 unit of sugar by only sacrificing 1 unit of corn at home, the U.S. will never be willing to pay more than 1 unit of corn in the world market.

This reasoning suggests that the domestic terms of trade serve as the maximum limits for the international terms of trade. Any international exchange ratio that falls within these limits, including the domestic exchange ratios themselves, is a possible outcome for the international terms of trade. For example, the international terms of trade might be the following:

$$\frac{\Delta S}{\Delta C} = -\frac{2S}{1C}$$

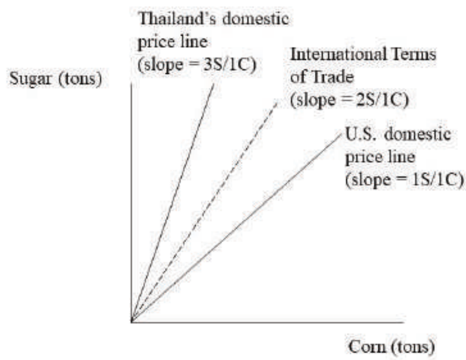
In other words, in the world market, Thailand must pay 2 units of sugar for 1 unit of corn. This outcome is possible because the world price of corn lies in between the domestic prices (i.e., between 1 and 3 units of sugar). Similarly, taking the reciprocal allows us to consider the world price of a unit of sugar.

$$\frac{\Delta C}{\Delta S} = -\frac{1/2C}{1S}$$

That is, 1 unit of sugar costs the U.S. 1/2 units of corn. The world price of sugar also lies between the domestic prices (i.e., 1/3 and 1 units of corn).

To visualize these limits to the terms of trade and this additional possible outcome for the international terms of trade, consider Figure 19.2.

Figure 19.2: The Limits to the International Terms of Trade



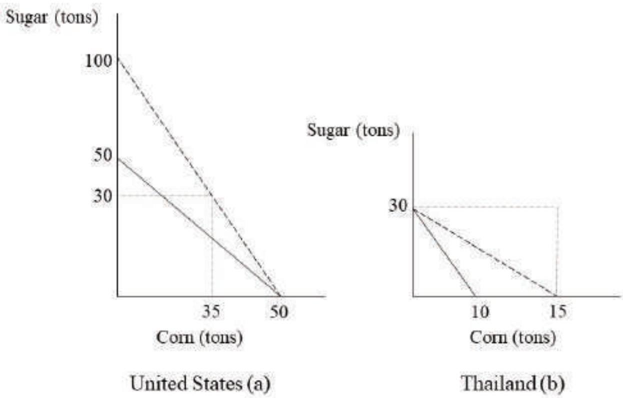
In Figure 19.2, the domestic marginal opportunity costs are represented as straight lines stemming from the origin. The negative signs of the slopes have been eliminated, but a tradeoff between the two commodities in each nation is implied. Because the domestic marginal opportunity costs are constant in each nation by assumption, these lines have constant slopes. The international terms of trade line must lie somewhere in between these extreme possibilities (although the extremes cannot be ruled out as possibilities for the established international terms of trade). If the international terms of trade line coincides with one of the domestic price lines, then that nation does not benefit from trade, and the other nation enjoys all the gains from trade. In this case, the international terms of trade line has a slope of 2 and thus it lies in between Thailand's domestic marginal opportunity cost line, which has a slope of 3, and the U.S. domestic marginal opportunity cost line, which has a slope of 1.

The final step in this section is to determine what effect these specific terms of trade have on the consumption possibilities of Thailand and the U.S. Each nation is engaged in complete specialization in the production of the commodity in which it has a comparative advantage. Given the international terms of trade that have been established, each nation can export a part of its production in exchange for a definite amount of the other nation's commodity. For example, the U.S. could hypothetically trade all 50 tons of corn for 100 tons of sugar at these specific terms of trade. Multiplying the numerator and denominator of the international terms of trade fraction by 50, the calculation may be completed as follows:

$$\frac{\Delta S}{\Delta C} = -\frac{2S}{1C} = -\frac{100S}{50C}$$

Because the U.S. can hypothetically reach this point of 100 tons of sugar as well as any other point along the straight line connecting these two points, we can add a **trading possibilities frontier (TPF)** to our earlier graph from Figure 19.1 (a). This graph is shown in Figure 19.3 (a).

Figure 19.3: Expanded Consumption Possibilities: The Trading Possibilities Frontier



Similarly, Thailand can hypothetically export all its sugar for 15 tons of corn. Following a similar approach as the one used to obtain the TPF of the U.S., we can arrive at this result by multiplying the numerator and denominator of the international terms of trade fraction by 30:

$$\frac{\Delta C}{\Delta S} = -\frac{1/2C}{1S} = -\frac{15C}{30S}$$

Since Thailand can trade with the U.S. to reach any point on the straight line connecting these two points, it also has a trading possibilities frontier (TPF). Thailand's TPF is represented in Figure 19.3 (b).

The major lesson from this analysis is that each nation can reach consumption combinations of the two commodities that lie outside their production possibilities frontiers. Neither nation has acquired more resources or more advanced production technologies, and yet each nation is able to consume more than before. Specialization and trade are what make this extraordinary result possible.

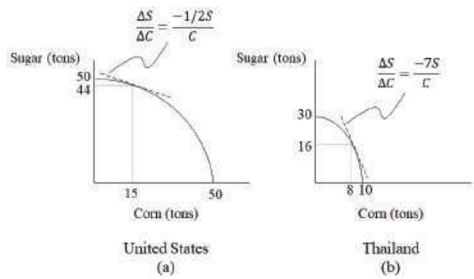
The careful reader might notice one little problem with the analysis. The maximum amount of sugar production in Thailand with complete specialization is 30 tons of sugar. Therefore, it is impossible for the U.S. to trade all its corn in exchange for 100 tons of sugar. Nevertheless, if it trades 15 tons of corn for 30 units of sugar, then it achieves a much higher level of consumption than what it can achieve without trade. The same holds true for Thailand. Aside from the complication of limited production in Thailand, each nation can achieve consumption possibilities that lie in between their PPFs and their TPFs. These results, therefore, suggest that international trade leads to mutual gains, and it is, therefore, the main reason why neoclassical economists argue so forcefully for free trade in the world market.

The Modern Theory of Comparative Advantage: The Case of Increasing Opportunity Cost

Now that we have a clear understanding of how nations may enjoy mutual gains from trade in the case of constant marginal opportunity costs, we can consider the case of increasing marginal opportunity costs. That is, we will assume that nations have production possibilities frontiers like those we considered in Chapter 2. Again, the underlying assumption that gives rise to this pattern of increasing opportunity cost is the assumption of heterogeneous resources.

Let's assume that the U.S. and Thailand have the PPFs shown in Figure 19.4., and that each nation is producing at the points identified on their PPFs in the figure.

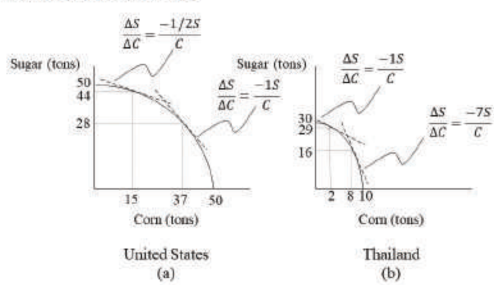
Figure 19.4: PPFs for Two Nations Facing Increasing Marginal Opportunity Costs



In this example, the U.S. is producing 15 tons of corn and 44 tons of sugar in autarky. **Autarky** refers to a closed economy, or an economy that does not engage in trade relationships. That is, it neither imports nor exports commodities. The slope of the tangent line drawn to the PPF at that point tells us the marginal opportunity cost of corn production. Let's assume that this slope is -0.5S/C. That is, the opportunity cost of producing 1 unit of corn is 1/2 unit of sugar in the U.S. as shown in Figure 19.4 (a). Similarly, Thailand is producing 8 tons of corn and 16 tons of sugar. The marginal opportunity cost of producing corn is also determined by the tangent line that just touches the curve at that point. Let's assume that the slope is -7S/1C. That is, the opportunity cost of producing 1 ton of corn in Thailand is 7 tons of sugar as shown in Figure 19.4 (b). It should be clear that the U.S. has a comparative advantage in corn production. Thailand would then have a corresponding comparative advantage in sugar production.

As before, each nation will specialize in the commodity in which it has a comparative advantage. That is, the U.S. will begin shifting resources towards corn production, and Thailand will begin shifting resources towards sugar production. Something interesting happens in this case, however, that is very different from the case of constant opportunity cost that we considered before. As the U.S. begins producing more corn, the marginal opportunity cost of corn production begins to rise. This increase occurs because less suitable resources must be increasingly relied upon as the best resources for sugar production are transferred to corn production. At the same time, Thailand's increased production of sugar causes the marginal opportunity cost of sugar production to rise for the same reason. Eventually, the marginal opportunity cost of corn production in each nation will be the same. That is, the domestic terms of trade will be exactly the same in each nation. At this point, the specialization ceases because neither nation will have anything to gain from further specialization. This situation is depicted in Figure 19.5.

Figure 19.5: Partial Specialization in the Case of Increasing Marginal Opportunity Cost

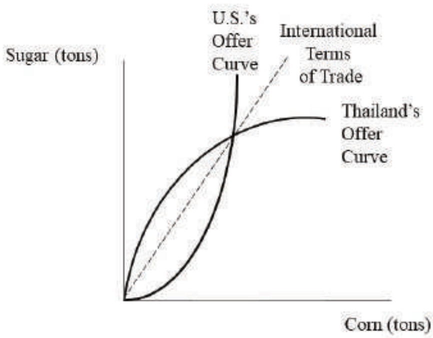


As Figure 19.5 shows, the two nations cease to increase their degrees of specialization once the domestic marginal opportunity costs in each nation are equal to 1 ton of sugar for each ton of corn. That is, once the slopes of the PPFs in each nation are equal to -1S/C, then the domestic prices in each nation are the same. The interesting result, in this case, is that each nation only pursues **partial specialization**. That is, each nation will not completely specialize in the production of the commodity in which it has a comparative advantage because its comparative advantage would then become a comparative *disadvantage*.

Once the marginal opportunity costs of the two nations have become equal, the limits to the terms of trade become identical for the two nations. Therefore, the international terms of trade must be the same as the domestic terms of trade once partial specialization is achieved. Unlike in the case of constant opportunity cost, the case of increasing opportunity cost leads to a unique outcome for the international terms of trade. Each nation partially specializes in one commodity and then exports some of that commodity in exchange for the commodity in which the other nation partially specializes.

As it turns out, we have only discussed part of the story. Demand also plays a role in determining the equilibrium terms of trade. To provide the complete picture, we would need an additional analytical tool to represent consumer preferences in each nation. Let's just move to a graph that will allow us to see how both supply and demand contribute to the determination of the equilibrium terms of trade. Figure 19.6 shows what are referred to as **offer curves** in mainstream trade theory, and it has its roots in the theory of international prices developed by John Stuart Mill.[1]

Figure 19.6: The Equilibrium International Terms of Trade in the Case of Increasing Marginal Opportunity Cost



The offer curve of the U.S. in Figure 19.6 shows how much corn the U.S. is willing to supply to Thailand for a given quantity of sugar. As the U.S. specializes more in corn, the marginal opportunity cost of corn production rises, which contributes to the rapid increase in the quantity of sugar required in exchange. Consumer preferences also play a role and so the offer curve reflects the impact of both production cost and consumer demand. Thailand also has an offer curve. Its offer curve shows how much sugar it is willing to supply in exchange for a given quantity of corn from the U.S. Its requirement of corn increases rapidly as well, as its production of sugar rises, as shown in Figure 19.6, which is also due to the rising marginal opportunity cost of sugar and the strength of demand.

The graph in Figure 19.6 shows quite clearly that the offer curves of the two nations intersect at a precise point. This intersection suggests that the U.S. wishes to export exactly as much corn as Thailand wishes to import and that Thailand wishes to export exactly as much sugar as the U.S. wishes to import. In other words, the quantities supplied and demanded of each commodity are equal in the world market. The result is equilibrium in the international marketplace. What guarantees this result? Let's assume that each nation must balance its international payments. That is, the value of each nation's exports must equal its imports. In other words, the following result must hold for each nation:

$$P_C C = P_S S$$

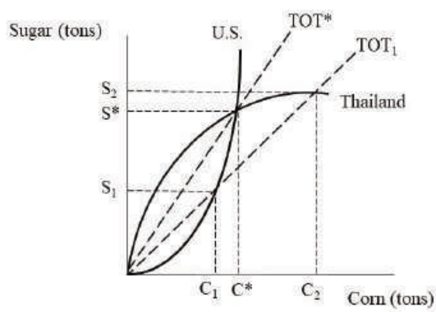
In the above equation, P_C represents the price of corn, C represents the quantity of corn, P_S represents the price of sugar, and S represents the quantity of sugar. Hence, the value of corn exported must equal the value of sugar imported in the U.S., and similarly, the value of corn imported must equal the value of sugar exported from Thailand. If we rearrange the equation, we obtain the following result:

$$\frac{S}{C} = \frac{P_C}{P_S}$$

This equation shows that the price of corn relative to the price of sugar (i.e., the relative price ratio) may be represented as the slope of a ray drawn from the origin, which may be calculated as the ratio of S to C at any point along that ray. If the ray becomes steeper, then the relative price of corn has risen. If the ray becomes flatter, then the relative price of corn has fallen. In addition, the nation's trade is balanced if it operates on this ray.

Now suppose that the international terms of trade are given by the line TOT_1 in Figure 19.7.

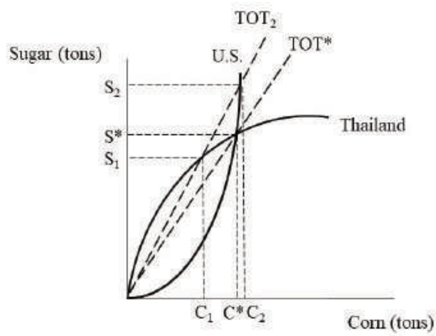
Figure 19.7: The Adjustment to Equilibrium in the Case of an Excess Demand for Corn and an Excess Supply of Sugar



In this case, if the U.S. produces and offers C_1 amount of corn in the market in exchange for S_1 amount of sugar and succeeds in doing so, then its trade will be balanced. The problem, however, is that Thailand will offer S_2 amount of sugar and wish to import corn of C_2. In other words, a large excess demand for corn will exist. Similarly, a large excess supply of sugar will exist. As a result, the relative price of corn will rise, and the relative price of sugar will decline. As the relative price of corn rises, the terms of trade line becomes steeper and moves in the direction of the equilibrium terms of trade line (TOT*) at the intersection of the two offer curves.

Next suppose that the international terms of trade are given by the line TOT_2 in Figure 19.8.

Figure 19.8: The Adjustment to Equilibrium in the Case of an Excess Demand for Sugar and an Excess Supply of Corn



In this case, Thailand produces and offers S_1 amount of sugar and wishes to import C_1 of corn. If it succeeds then its trade will be balanced. The problem this time is that the U.S. is offering a great deal more corn (C_2) and wishes to buy S_2 tons of sugar, which is much more than Thailand wishes to export. As a result, an excess demand for sugar exists and an excess supply of corn exists in the world market. As a result, the relative price of sugar will rise and the relative price of corn will fall. As the relative prices change, the ray from the origin, which represents the international terms of trade, becomes flatter. It is only when the ray passes through the point of intersection of the two offer curves that each nation is balancing its trade, which is reflected in the fact that each nation is operating on the terms of trade line and its offer curve.

Trade has become balanced due to international prices responding to the competitive pressures of supply and demand in the world market.

Criticisms of the Neoclassical Theory of Comparative Advantage

The theory of comparative advantage remains the foundation of the neoclassical theory of international trade to this day. It is more than just a theory of how nations engage in trade with one another. It is also an ideological defense of **free trade**. That is, economists reject any kind of government restrictions on international trade on the grounds that such restrictions undermine the potential for the mutual gains from trade to be realized. Because the theory of comparative advantage is the foundation of the free trade theory that most economists advocate, it is important to consider heterodox critiques of the theory to determine whether they have any merit.

One argument that is made in favor of free trade is that a nation's exports generate sufficient revenue to finance the purchases of imports. Hence, trade should be balanced. It is important to note, however, that just because a nation recently sold commodities does not mean that it must immediately buy commodities from another nation. A trade surplus may persist if the nation does not purchase imported commodities. Meanwhile the other nation has purchased commodities but is unable to sell. It will, therefore, experience a trade deficit. If it finances its purchases using foreign currency reserves and those reserves begin to run out, then the result may be a currency crisis. The argument that imports can always be financed by exports is rather similar to the argument that whenever commodities are produced, enough income is paid to the factors of production to ensure the purchase of those commodities. We identified this claim in Chapter 13 as Say's Law of Markets. The problem, of course, is that just because incomes have been received does not guarantee that they will be spent any time soon. If enough income is saved, then the result is a glut of commodities. In both examples, gaps between the time income is received and the time it is spent create the possibility of economic crises.

The problems with mainstream trade theory go far beyond this specific issue and relate to the assumptions underlying the theory of comparative advantage. In his book *Free Trade Doesn't Work* (2009), Ian Fletcher provides a detailed account of eight hidden assumptions that form the basis of the theory.[2] Fletcher argues that exposing the assumptions reveals the flaws inherent in the theory. This section and the next section provide a brief overview of Fletcher's insights.[3]

The *first hidden assumption* that Fletcher identifies is the assumption that trade is sustainable.[4] To finance its imports, a nation may not be exporting commodities but rather assets.[5] If it sells bonds, for example, then its debts may accumulate to the point where the future interest payments associated with that debt interfere with future capital investments and/or consumption. A nation may also specialize in the production of a non-renewable natural resource because it has a comparative advantage in its production.[6] Mutual gains from trade might exist in the short run, but in the long run, its exports will deplete its natural resource base to the point where it might experience a massive economic crisis. Fletcher offers Nauru as an example of an island nation that exported large quantities of guano, which is used for manufacturing fertilizer, from 1908 to 2002.[7] The economy boomed in the 1960s and 1970s. When the guano ran out, the economy collapsed. Middle Eastern nations that are heavily dependent on exports of petroleum might experience similar problems of economic dislocation as their oil reserves are depleted.[8]

The *second hidden assumption* that Fletcher identifies is the assumption that no externalities exist.[9] Ideally, the prices that emerge in the world market should reflect all costs and benefits associated with the production and consumption of the commodities traded. If third parties are affected by the production or consumption of the products, either positively or negatively, then positive or negative externalities exist and the commodities are either under-produced or over-produced, as discussed in Chapter 3. For example, if a production process leads to air pollution, then efficiency requires that the commodity produced be priced high enough to cover the costs to third parties who are then compensated for the harm done to them. If this corrective action is not taken, then the commodity will be produced in quantities that are too large, and the price will be too low.

The *third hidden assumption* is that factors of production are easily transferred from one industry to another.[10] Consider our earlier example involving constant opportunity costs in which the U.S. completely specialized in corn production and Thailand completely specialized in sugar production. Complete specialization requires a nation to shift all its resources away from the production of the commodity in which it has a comparative disadvantage. The problem is that transferring resources may not be an easy task to accomplish. Factories need to be converted to an entirely new type of production. Workers who may be considerably skilled at producing one type of commodity must now produce an entirely different commodity. The training involved may be expensive and time-consuming. The result very well may be unemployment of labor and production capacity that is not used for a long time. The theory of comparative advantage is entirely static. That is, the passage of time is not taken into account. Therefore, this problem is not obvious when considering movements along a production possibilities frontier, but the problem will be very real for workers and factory owners when cheaper imports begin to enter their nation.

The *fourth hidden assumption* is that international trade unambiguously raises the welfare of everyone in the nation.[11] After all, each nation can import the commodity it desires at a lower price. In our constant opportunity cost example, the U.S. acquires sugar at a lower price than it could domestically, and Thailand acquires corn at a lower price than it could domestically. We also saw that both nations enjoyed trading possibilities frontiers that lie beyond their production possibilities frontiers. The problem, however, is that with some industries contracting and other industries expanding, the gains are not equally shared. In fact, some workers and capital owners will lose income as a result of trade. A net gain for society as a whole does occur, however, because the winners gain more than the losers lose. According to the **Kaldor-Hicks criterion**, as long as the winners can hypothetically compensate the losers, the situation represents an improvement (even if the compensation does not actually occur). Nevertheless, the welfare gain for the nation may not be shared by all, and so this criticism is worth considering when evaluating the comparative advantage model.

The *fifth hidden assumption* is that capital is not internationally mobile.[12] If capital is able to leave the nation and seek more productive opportunities in other nations, then the comparative advantage model does not apply. Remember that the nation's resource stocks determine the position of its PPF. If a nation begins to lose capital as trade begins, then its PPF will shift inwards. None of the results we discussed earlier would apply in such a situation. When Ricardo first developed the theory of comparative advantage, the assumption of capital immobility was more realistic. In the modern world, capital is highly mobile and so the argument is that the comparative advantage model does not apply very well today.

The *sixth hidden assumption* is that short-term efficiency is compatible with long-term economic growth.[13] As a static model, the theory of comparative advantage has nothing to say regarding the long-term growth prospects of the nation. Indeed, it is possible that a nation that chooses to specialize in the production of a commodity in which it has a comparative advantage might become permanently stuck producing a commodity that does not promote long-term economic growth. The original example that Ricardo used involved Britain's specialization in textile production and Portugal's specialization in wine production. As Fletcher points out, the British textile industry promoted technological change with the development of steam engines and sophisticated machine tools.[14] Wine, on the other hand, was produced using traditional methods that did not encourage innovation or productivity improvements.[15] Although Portugal may have benefited from specializing in wine production at the time, it is now the poorest nation in Western Europe.[16] Its goal of static efficiency appears to have been achieved at the expense of its long term growth prospects.[17]

The *seventh hidden assumption* is that trade does not induce adverse productivity growth abroad.[18] It is possible that trade might promote foreign industries to the point where their comparative advantage begins to change. In that case, the commodities that the foreign nation previously supplied are no longer supplied. The example that Fletcher gives is that of Japan in the 1950s and 1960s.[19] As the Japanese economy boomed, it transitioned from providing the U.S. with cheap manufactured commodities to those that required more sophisticated manufacturing processes.[20] This problem is not as serious as some of the others mentioned because the nations still gain from trade. The argument appears to be that the importing nation might not gain as much as previously.

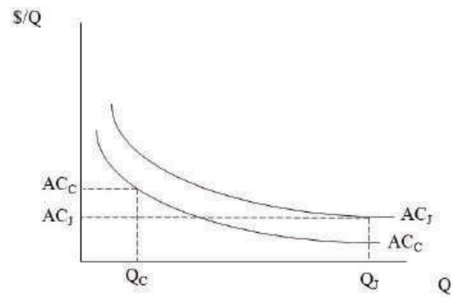
Developments in New Trade Theory: The Potential Impact of Economies of Scale

Ian Fletcher's *eighth hidden assumption* is that no economies of scale exist in production.[21] This assumption leads to a discussion of what has become known as New Trade Theory. One of the major contributions to New Trade Theory was made by Ralph Gomory and William Baumol in their book *Global Trade and Conflicting National Interests* (2000).[22] In their book, Gomory and Baumol assume that economies of scale in production do exist.[23] Therefore, they depart from the traditional assumption implicit in the theory of comparative advantage that no scale economies exist.

To understand how scale economies lead to different theoretical results, we need to understand what Gomory and Baumol mean by "retainable industries." **Retainable industries** are industries that a nation captures and retains due to a cost advantage stemming from economies of scale.[24] To capture such industries, the nation must be the first to achieve a large enough volume that such low per unit production costs cannot be easily matched by competitors in other nations.[25] What is interesting about retainable industries is that a nation might capture one even if another nation would turn out to be a superior producer if it managed to reach the same large volume of production as its competitor. The reason that the superior producer fails to do so is that scale economies serve as a barrier to entry, preventing the potential competitor from entering the market. Fletcher refers to this possibility as the **lockout phenomenon**.[26]

Figure 19.9 provides an example involving Japan and China where each nation enjoys economies of scale in production as reflected in the downward slope of their average cost (AC) curves.[27]

Figure 19.9: Economies of Scale in Retainable Industries



It should be clear from Figure 19.9 that China is the superior producer. At any level of output, China has a per unit production cost that is below Japan's per unit cost. Because Japan entered the industry much earlier, however, its larger production level of Q_J allows it to produce at a much lower per unit cost than China, which produces an output level of Q_C. Clearly, if China expanded production to match Japan's output, then its unit cost would be lower and it would capture the industry, but because its unit cost is higher, firms will not be able to compete and its superior efficiency will not be realized. In this case, the industry is a retainable industry for Japan, and the world economy loses out on an opportunity to produce the product at a lower unit cost. That is, free trade may not produce the best result for the world economy because many efficient producers will be locked out.

Fletcher draws upon this theory to explain why

Bangladesh exports many T-shirts but few soccer balls while Pakistan exports many soccer balls but few T-shirts.[28] That is, economies of scale allowed these nations to capture specific industries, and those nations that achieve large production volumes first are the most likely to retain those industries.

It might seem that a nation's best trade strategy in a world of scale economies is to capture as many retainable industries as possible by increasing production rapidly to drive down unit costs.[29] Fletcher explains that a nation should pursue such a strategy but only up to a point.[30] That is, if a home nation captures too many retainable industries, then it might actually prevent other nations from producing and exporting commodities to the point where it misses out on the superior efficiency of its competitors.[31] It might also leave other nations impoverished and unable to buy the exports of the home nation.[32] Furthermore, the home nation might find its own resources to be spread too thinly.[33] To see how a nation might go too far, consider the graph in Figure 19.10.[34]

Figure 19.10: The Effect on GDP of Capturing Retainable Industries

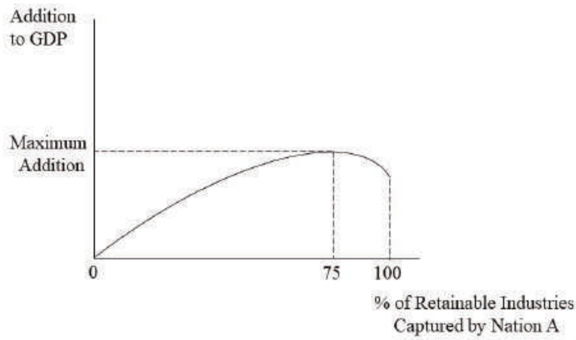


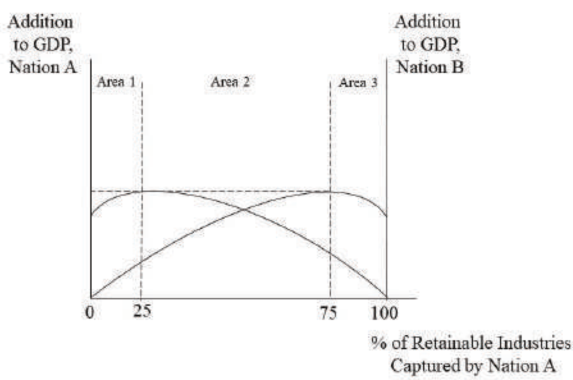
Figure 19.10 shows what happens to a nation's GDP as it captures a larger percentage of the world's retainable industries. Its GDP increases with the capture of more industries but only up to a point. Beyond that point, its GDP begins to fall for the reasons stated.

The analysis has a startling implication for trade theory. It suggests that trade is not always a positive sum game.[35] A **positive sum game** refers to a situation in which all players may gain without anyone losing. The theory of comparative advantage suggests that competitive interaction in the international marketplace is a positive sum game. The economic pie is made larger as both nations experience expanded consumption possibilities. The potential for a negative impact on income distribution is not explicitly captured in the model and so it appears to be a positive sum game. A **zero-sum**

game, on the other hand, occurs when one player may gain but only at the expense of another player. The suggestion of New Trade Theory is that international trade is sometimes a zero-sum game in which one nation gains at another nation's expense. This perspective directly contradicts what mainstream economists have long argued about the mutual benefits from international exchange.

To see how the assumptions of New Trade Theory generate this interesting result, consider Figure 19.11.[36]

Figure 19.11: The Distribution of Gains from Trade with the Capture of Retainable Industries



In Figure 19.11, we see how each nation's GDP changes as it captures or loses retainable industries. As nation A increases its share of retainable industries in Area 1, its GDP increases. Notice that nation B also experiences an increase in GDP as it loses retainable industries. The reason is that nation B would be overreaching if it acquired such a large percentage of retainable industries. Giving up some of these industries to nation A actually benefits nation B. Both nations gain from trade, and this result is consistent with what neoclassical economists have long argued about the mutual gains from international trade.[37]

We discover a similar result in Area 3. In Area 3, nation A experiences a loss of GDP as a result of overreaching. It has captured so many retainable industries from nation B that its GDP begins to fall. Therefore, Nation A will prefer to capture fewer industries. Nation B, on the other hand, will prefer to capture more retainable industries because it has not yet captured many. In doing so, its GDP will rise so it will happily acquire the retainable industries that Nation A is willing to sacrifice. We see another example of a positive sum game and of mutual gains from trade, consistent with the traditional neoclassical conclusions about free trade.

It is in Area 2 where we see a very different result. In Area 2, Nation A gains from capturing more retainable industries. Nation B, however, also wants retainable industries and so any gains in GDP for Nation A must come at the expense of Nation B. Similarly, any gains in GDP for Nation B must come at the expense of Nation A in this region. This area is characterized by mutual conflict rather than mutual benefit.[38] A theory of international trade that suggests trade is a zero-sum game is one that contrasts very sharply with conventional trade theory.

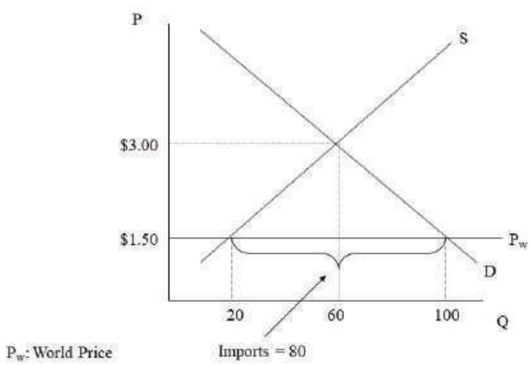
This New Trade Theory suggests that foreign productivity growth might lead to a loss of GDP, sending the gains from trade into negative territory.[39] It also suggests that the consequences of trade are rather complicated, sometimes leading to mutual benefit and sometimes leading to conflict.[40] Fletcher's policy conclusion is that free trade should be the rule in Ricardian industries in which no economies of scale are present.[41] Because of the existence of retainable industries, however, he favors something called **rational protectionism**.[42] That is, free trade should be the rule in Ricardian industries, but protectionism should be used in the case of retainable industries. For example, subsidies should be given to industries that are relatively new but in which large scale economies exist. Such subsidies will allow firms to break into industries in which the low unit costs of foreign rivals are likely to keep them out.[43] Infant industry tariffs are another option to protect domestic industries from foreign competition so that they have an opportunity to grow and take advantage of scale economies.[44] Infant industry tariffs are taxes on imported commodities aimed at protecting fledgling domestic industries.

Although he advocates rational protectionism, Fletcher does acknowledge the difficulties associated with such a policy stance. One of the main challenges is knowing which industries to target.[45] Another difficulty is that many modern corporations have become multinational in scope. It may be difficult to identify the home nation in which case it is not clear how such firms would even fit into this analysis.[46]

The Neoclassical Critique of Protectionist Policies: Import Tariffs and Quotas

In addition to the neoclassical theory of comparative advantage, mainstream economists also rely on partial equilibrium analyses to defend the claim that protectionist policies reduce social welfare by undermining efficiency. To understand why, let's consider the case of a small nation that has a domestic supply and demand for a specific good. Furthermore, let's assume that it can import all it wants at the world market price (P_W). This situation is depicted in Figure 19.12.

Figure 19.12: International Trade and the Supply and Demand Model

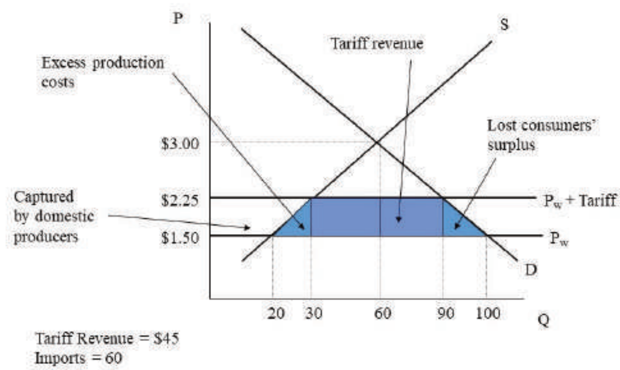


In this market, the equilibrium autarky price is $3.00 per unit, and the equilibrium quantity exchanged is 60 units. Because trade is permitted, however, and the world price is only $1.50 per unit, the domestic producers will only produce and sell 20 units. At that price, the remaining 80 units of the 100 units demanded will be imported. Free trade, therefore, allows the small nation to consume a larger quantity of the good and at a lower price than in the case of autarky. This result is basically consistent with the conclusion of comparative advantage theory.

The reader should also recall that the welfare of the consumers in the market can be measured using consumers' surplus. In Figure 19.12, consumers' surplus is equal to the area below the market demand curve and above the world price line at $1.50. In this case, no producers' surplus exists because the world supply curve coincides with the world price line and so the price received by producers is just enough to cover production costs. Therefore, the large triangle representing consumers' surplus also represents the total surplus or total welfare of all market participants.

The next question that we might consider is what will happen if the government decides to impose a tariff on imports. An **import tariff** is simply a tax on imported goods. In this case, the tariff is assumed to take the form of a tax per unit of physical product. For example, if the tariff is $0.75 per unit and the world price is $1.50 per unit, then the producers in the rest of the world must simply add $0.75 to the world price. The new world price then becomes $2.25 per unit. This situation is depicted in Figure 19.13.

Figure 19.13: The Economic Consequences of an Import Tariff



In Figure 19.13, we see that the effect of the higher price is to reduce the quantity demanded of the product domestically from 100 units to 90 units. In addition, the higher price encourages domestic producers to produce and sell 30 units of the good rather than the 20 units previously sold. Imports thus fall to 60 units, which is the difference between the quantity demanded and the quantity domestically supplied.

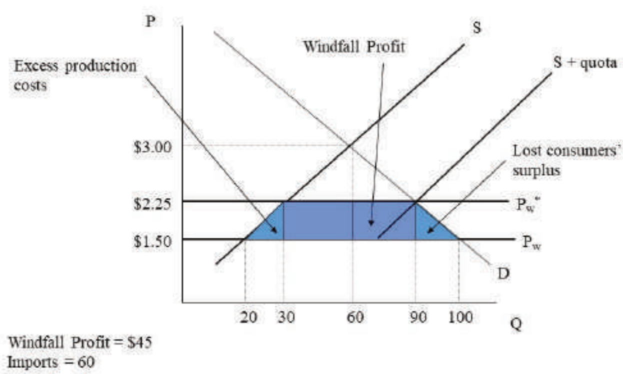
In terms of the impact on social welfare, we need to consider what happens to consumers' surplus due to the tariff. In Figure 19.13, all the areas labeled represent lost consumers' surplus due to the higher world price. One of these regions, however, is captured by domestic producers in the form of producers' surplus, which they receive now that they are able to charge a higher price as well. This gain is the portion above the domestic supply curve (and between the old and new prices). The portion that is below the domestic supply curve from units 20 to

30 (and above the old price) represents excess production costs that the higher world price allows the domestic producers to incur. In other words, the higher world price allows the domestic producers to produce inefficiently. Because it does not represent producers' surplus and it is lost as consumers' surplus, it therefore represents a deadweight loss to society.

An additional deadweight loss to society that results from the tariff derives from the reduction in quantity demanded that occurs when the world price rises. Because consumers consume 10 fewer units as a result of the tariff, it is not possible for any consumers' surplus to be realized on these units. Furthermore, since the units are not produced, producers cannot capture these gains, and the welfare simply vanishes. The final portion that must be considered is the rectangle in the middle, which represents tariff revenue. The base of the rectangle represents the quantity of imports, which in this case is 60 units. The height represents the tariff per unit of $0.75. If we multiply the tariff per unit times the quantity of imports, we obtain the total tariff revenue of $45 (= $0.75 per unit times 60 units). This portion of the lost consumers' surplus is captured by the government and so it is not a deadweight loss. The government has the power to use this revenue to provide services for the nation's citizens and so it does not necessarily lead to a shrinking of the economic pie. Overall, however, the economic pie shrinks due to the excess production costs that protected firms are able to incur and the reduction in quantity demanded resulting from the higher world price.

One final protectionist policy that we might consider is an **import quota**. An import quota is simply a quantitative limit on the amount of a good that may be imported into a nation. For example, the government might set an import quota of 60 units. Therefore, the maximum quantity that may be imported of the product is 60 units. What the quota means is that the total supply of the product will now be equal to the U.S. domestic supply plus the quota amount. That is, the domestic supply curve will shift to the right by the amount of the quota, as shown in Figure 19.14.

Figure 19.14: The Economic Consequences of an Import Quota



The intersection between the new supply curve and the demand curve gives us an equilibrium price of $2.25 and an equilibrium quantity exchanged of 90 units. The result in this case looks very much like the result we obtained in the case of an import tariff. In fact, the new world price and the quantity of imports are the same as in the case of the tariff. The welfare effects are also similar. We see that more domestic producers produce the commodity at a higher cost thanks to the restriction on imports, which makes possible the higher price. We also see that the quantity demanded drops leading to a lower consumers' surplus. Both of these effects contribute to the deadweight loss of the policy.

As before, the domestic producers do gain producers' surplus at the expense of consumers. The one consequence of the import quota that makes this case rather different from the import tariff is the portion of lost consumers' surplus represented with the rectangle. Because the government does not collect a tax but simply limits the quantity of imports, who collects this extra revenue from the sale of the product? As Robert Carbaugh explains, the answer depends on whether foreign exporters or domestic importers have greater market power.[47] For example, if the foreign exporters collude, then they can demand a price of $2.25 per unit from domestic importers. The exporters will then gain the entire amount as a windfall profit. On the other hand, if the importers collude and refuse to pay any price higher than $1.50 per unit, then when they turn around and sell at $2.25 per unit to the consumers, the importers will gain the entire amount as a windfall profit. A final possibility that is worth considering is that the government sells import licenses to domestic importers that grants them the legal right to import the good.[48] If it charges the maximum possible fee for these licenses, then it may end up capturing the entire windfall profit itself.[49] In this last case, the consequences of the import quota are identical to the consequences of the import tariff since the government captures the rectangle of lost consumers' surplus. The major point to notice, of course, is that in both the import tariff and the import quota cases, overall social welfare is reduced due to the deadweight losses that result from the higher world price and the subsequent responses of domestic producers and consumers.

The Challenge of Dependency Theory

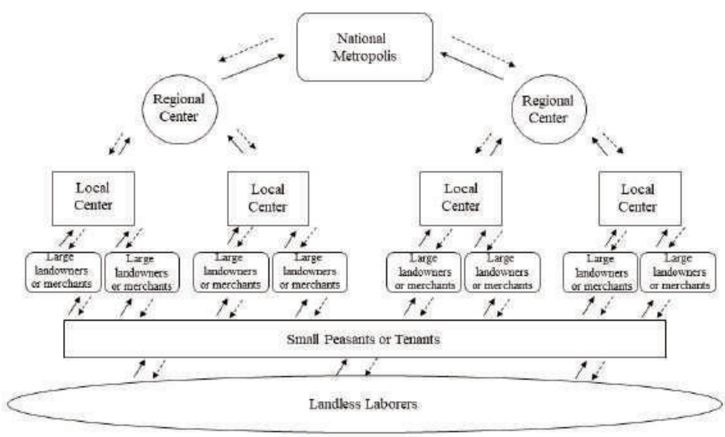
Another radical theory of the world capitalist economy is known as **dependency theory**. According to this theory which first appeared in the 1960s, the world capitalist economy consists of rich, capitalist nation-states that exploit poor, developing nation-states by taking advantage of their abundant natural resources and cheap labor-power. The exploitative relationship that results creates a situation in which the poor nations become economically dependent on the rich nations and also become stuck in a chronic state of **underdevelopment** as a consequence of that relationship.

Andre Gunder Frank is one of the most important contributors to dependency theory. His *Capitalism and Underdevelopment in Latin America* (1969) has been highly influential in developing this theoretical framework. This section mainly concentrates on his perspective. According to Frank, because power and resources are unequally distributed throughout the world economy, some nations develop more rapidly than others.[50]The result has been underdevelopment in less developed nations and rapid economic growth in advanced industrialized nations. The theory asserts that the world capitalist economy is divided into a series of networks that link a few capitalist metropolises to many satellite nations. Sometimes dependency theorists refer to the **center** and the **periphery** rather than to the **metropolis** and its **satellites**. The metropolis acquires raw materials at low cost from the periphery and uses them to produce finished commodities in the center. The commodities are sometimes exported to the periphery, which allows the center to continue its appropriation of surplus production in the satellite nations.[51] The Marxist roots of the theory are apparent in the emphasis on surplus production and appropriation. Indeed, Frank's thesis is that the contradictions in the world capitalist economy have allowed metropolitan centers to appropriate the surpluses of peripheral satellites to the benefit of the former and to the detriment of the latter.[52]

Monopoly capital plays an important role in Frank's theory. Indeed, monopoly characterizes the world capitalist economy in his view.[53] Frank asserts that the world capitalist system and the periphery have possessed an extremely monopolistic structure throughout their history.[54] The metropolis uses monopoly power in markets in which it is a seller, and monopsony power in markets in which it is a buyer, to extract the surplus product from its satellites. It then refuses to invest in the satellites, which stunts their growth.[55] Over time, the satellites become increasingly dependent on the metropolis, which creates distortions in the satellites' economies.[56] A local ruling class in each satellite, referred to as the **lumpenbourgeoisie**, ensures that this system of exploitation persists.[57] It captures a part of the surplus as it is transferred in an upward direction and so the system serves the interests of the lumpenbourgeoisie.

In Frank's theory, this network of metropolises and satellites contains several levels of surplus appropriation. Chains of metropolis-satellite connections make up the global capitalist order with each metropolis extracting surpluses from its satellites and with a metropolis sometimes serving a higher-order metropolis.[58] The chains also operate within nations and between them, creating “an extended continuum of exploitative relationships.”[59] Frank describes how the national metropolises appropriate the surpluses of regional centers and how the chain of surplus appropriation continues down the chain to local centers, then to large landowners or merchants, then to small peasants or tenants, and sometimes even to landless laborers.[60] Figure 19.15 shows how these different levels of surplus appropriation relate to one another.

Figure 19.15: A Metropolis and Its Satellites



The world capitalist system consists of several national metropolises and each exercises its monopoly power. This position of power makes it possible for the metropolis to transfer surplus production to itself and away from the periphery through its control of pricing. The solid arrows and the dashed arrows in Figure 19.15 indicate that the flows traveling in an upward direction are larger than the flows traveling in the downward direction at each level. The differences between the upward flows and the downward flows represent the surplus.

The hierarchical structure depicted in Figure 19.15 does not exist within a competitive capitalist framework. That is, in competitive capitalism, workers confront capitalists but no complicated chain of metropolises and satellites will arise. The reason is that monopoly power is required for these networks to become established. Indeed, two kinds of monopoly give rise to these networks within the world capitalist economy.[61] The first kind is **monopolistic merchant's capital** where merchants purchase products from local producers and then export them.[62] These merchant capitalists are not directly involved in production.[63] The second type is **modern monopoly capital**, which is based on large-scale capitalist production and modern production technology.[64]

Frank's major point is that underdevelopment results from these interactions. Without investment in sectors that will promote the growth of employment and production in the satellites, the satellite nations will remain at a low level of development. Unfortunately for the periphery, it is the metropolis that decides how to use the large surpluses appropriated from the satellites and collected at the center. This situation poses long-term development problems for developing nations.

The development of capitalism on a global scale since the 1960s provides an abundance of evidence to test the theory. According to Felipe Antunes de Oliveira, "Frank's prediction that no real development was possible within capitalism was almost immediately challenged by the facts."[65] Economic growth in Latin America in the late 1960s and early 1970s and the rapid growth of East Asian countries in the 1980s and 1990s "seemed to provide further and conclusive evidence against Frank's dependency theory."[66] The burden then is on dependency theorists to explain these cases in a manner that is consistent with their overall approach to the world capitalist economy. Nevertheless, Antunes explains that the economic stagnation in the 1980s and more recent economic crises in Latin America imply that Frank's theory "may still capture a deeper truth about the limits to peripheral development."[67] Despite some challenging evidence, dependency theory provides us with a way of thinking about how aggregate surplus production is appropriated and redistributed throughout the world capitalist economy.

The Theory of Unequal Exchange

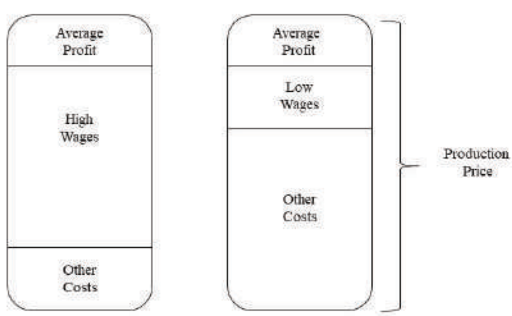
One final alternative radical theory of world trade that we will consider is Arghiri Emmanuel's theory of **unequal exchange**, which he developed in the early 1970s in his *Unequal Exchange: A Study of the Imperialism of Trade* (1972) . Emmanuel aimed to show that competitive interaction between rich and poor countries in the global marketplace could lead to a worsening terms of trade for poor countries. To obtain this result, Emmanuel built some key assumptions into his model. For example, he assumed that the world economy only consists of two nations. He also assumed that the profit rates in the rich and poor nations would equalize as a result of the free flow of global capital. If the profit rate of one nation exceeds the profit rate of another nation, then capital will flow to the nation with the highest profit rate, driving that profit rate down and pushing the other nation's profit rate up until equalization occurs.

At the same time, Emmanuel assumes that wage rates in the rich and poor nations differ significantly because labor is not mobile. As Itoh explains, capital is relatively mobile, but labor is relatively immobile in Emmanuel's framework due to immigration restrictions.[68] Howard and King also state that Emmanuel's theory assumes "a powerful tendency for the rate of profit to be equalized on a world scale, while there remain huge differences in both wage rates and rates of exploitation between advanced and backward countries."[69] As Emmanuel explains, he treats wages as the independent variable in his theoretical system.[70] Furthermore, he assumes that the rates of surplus value are "institutionally different" and not subject to competitive factor market equalization.[71] The fact that differences in rates of surplus value exist alongside wage differentials should make sense. If two nations have similar working day lengths but one nation has much lower wages, then the nation with much lower wages should have a higher rate of surplus value, other factors the same.

The persistent wage differentials are explained in terms of the immobility of labor, but what about the absolute levels of wages in the two nations? Emmanuel's assumption about differential wages stems from Marx's claim that the value of labor-power contains an "historical and moral element."[72] The reader should recall that the value of labor-power depends on cultural and social factors that are specific to time and place. What constitutes a required bundle of commodities in a rich nation will be very different from the required bundle of commodities in a poor nation. Generally, the workers in a rich nation will be accustomed to having more and better quality commodities available to them than workers in poor nations. Hence, the value of labor-power and the wage rate will be higher in the rich nation and lower in the poor nation.

One might suspect that the wage differentials will lead to different prices across nations for the same product. This result does not necessarily follow though. For example, Emmanuel considered the possibility that two capitalist nations produce the same commodities but that one nation has a higher wage level.[73] For a common rate of profit and for a single production price to emerge, the higher wage nation must have higher productivity so that its other costs are lower and its overall costs are the same across the two nations.[74] The situation is easier to visualize by considering the diagram in Figure 19.16.

Figure 19.16: A Case of Price and Profit Rate Equalization with Differential Wages



As Figure 19.16 shows, the higher wage nation has other production costs that are lower, which makes it possible for costs to be equal across the two nations. Indeed, the higher productivity is the cause of the higher wages. When adding the average profit to the equal costs, the production prices are the same. We thus see one example of how profit rates and production prices can be the same even when wages vary across nations.

We next consider an example in which two nations produce different commodities. As we learned in Chapter 8, Marx transformed values into production prices to explain how a general rate of profit is formed in the economy. Setting aside the criticism of Marx's procedure, which has given rise to the Transformation Problem debate, we follow Emmanuel in using Marx's procedure to equalize the rate of profit throughout the world economy. Let's consider only two countries: the United States and Mexico. Consistent with Emmanuel's approach, we will assume that one nation, in this case the United States, pays a higher wage to workers than the other nation (Mexico). Howard and King use a simple numerical example involving the same organic compositions of capital in two nations with one nation having a lower wage rate and a higher rate of surplus value.[75] Here we consider a similar example using basic algebra to illustrate the main findings of the theory.

For notational purposes, the high-wage United States will labeled nation A, and low-wage Mexico will be labeled nation B. The two nations produce different goods, but the organic composition of capital is the same in the two nations. The equal organic compositions of capital may be represented as follows:

$$\frac{C_A}{V_A} = \frac{C_B}{V_B}$$

In this equation, C and V refer to the constant capital and the variable capital in each nation. It is also assumed that wages are lower in Mexico, and the rate of surplus value is correspondingly higher in Mexico. These conditions are represented as follows:

$$V_A > V_B \; and \; \frac{S_B}{V_B} > \frac{S_A}{V_A}$$

In this inequality, S refers to the surplus value in each nation. We can now write the total value produced in each nation as the product of the physical quantity produced (q) in that nation and the value per unit (w) in that nation. The following equations thus hold:

$$q_A w_A = C_A + V_A + S_A$$

$$q_B w_B = C_B + V_B + S_B$$

The reader should recall that to obtain the rate of profit, we need to add the entire surplus value (S) in both nations and divide by the aggregate capital (C+V). Without subscripts, these aggregates refer to the combined sums for the two nations. That is,

$$r = \frac{S}{C+V}$$

The total production price (qp) in each nation may now be computed as the product of the output (q) in that nation and the per unit production price (p) in that nation. The total production price in each nation is then equal to the sum of the capital advanced and the average profit in that nation as follows:

$$q_A p_A = C_A + V_A + r(C_A + V_A)$$

$$q_B p_B = C_B + V_B + r(C_B + V_B)$$

Using this information, we can calculate the **value ratio** and the **production price ratio**. The value ratio is the ratio of the individual unit values (based on embodied labor) in the two nations. The price ratio is the ratio of the per unit production prices in the two nations. The ratios are calculated as follows:

Value ratio: $\frac{w_A}{w_B} = \frac{\frac{C_A+V_A+S_A}{q_A}}{\frac{C_B+V_B+S_B}{q_B}} = \frac{C_A+V_A+S_A}{C_B+V_B+S_B} \cdot \frac{q_B}{q_A}$

Production price ratio: $\frac{p_A}{p_B} = \frac{\frac{C_A+V_A+r(C_A+V_A)}{q_A}}{\frac{C_B+V_B+r(C_B+V_B)}{q_B}} = \frac{C_A+V_A+r(C_A+V_A)}{C_B+V_B+r(C_B+V_B)} \cdot \frac{q_B}{q_A}$

To give these equations meaning, suppose that we have the following information for the U.S. and Mexico:

Nation	Constant Capital	Variable Capital	Surplus Value	Quantity Qroduced (q)
United States (A)	600	300	200	130
Mexico (B)	300	150	300	110
Aggregate	900	450	500	

It should be clear that the organic compositions of capital are the same in the two nations (C/V = 2). It should also be clear that the rate of surplus value (S/V) is higher in Mexico where wages (V) are lower. Using our equations to calculate the value ratio, the rate of profit, and the production price ratio, we obtain the following:

Value ratio: $\frac{w_A}{w_B} = \frac{C_A+V_A+S_A}{C_B+V_B+S_B} \cdot \frac{q_B}{q_A} = \frac{1100}{750} \cdot \frac{110}{130} = 1.24$

Rate of profit: $r = \frac{S}{C+V} = \frac{500}{1350} = 37.04\%$

Production price ratio: $\frac{p_A}{p_B} = \frac{C_A+V_A+r(C_A+V_A)}{C_B+V_B+r(C_B+V_B)} \cdot \frac{q_B}{q_A} = \frac{1233.33}{616.67} \cdot \frac{110}{130} = 1.69$

The example shows that the production price ratio exceeds the value ratio. The implication is that the price of the commodity in the United States is relatively higher than what it should be given its relative value. Another way of stating the same result is to say that international competition causes the price of commodity A to be 1.69 units of commodity B. At the same time, if the commodities exchanged according to their values (i.e., labor content), then the price of commodity A would be 1.24 units of commodity B. The price of the U.S. commodity is thus higher than it would otherwise be. As Howard and King explain, the unequal exchange of commodities that results causes a transfer of value to the rich country, which worsens global inequality over time.[76] As the reader might recall, Marx showed how capitalist societies produce surplus value even when commodities exchange according to their values. He did acknowledge, however, that unequal exchanges can lead to the redistribution of value with one side to the exchange gaining and the other side losing. Emmanuel has shown how nations can be involved in such unequal exchanges leading to systematic value transfers from poor nations to rich nations.

A general claim may thus be stated and proven: If the production price ratio (p_A/p_B) exceeds the value ratio and two nations organic compositions of capital are the same, then the rate of surplus value must be higher in nation B. To prove this result, we proceed as follows:

$$\frac{p_A}{p_B} > \frac{w_A}{w_B} \Rightarrow \frac{C_A+V_A+r(C_A+V_A)}{C_B+V_B+r(C_B+V_B)} > \frac{C_A+V_A+S_A}{C_B+V_B+S_B}$$

With a bit of algebraic manipulation, this equation may then be rewritten as follows:

$$(\frac{C_B}{V_B}+1+\frac{S_B}{V_B})\cdot r(\frac{C_A}{V_A}+1)+\frac{S_B}{V_B}\cdot(\frac{C_A}{V_A}+1) > (\frac{C_A}{V_A}+1+\frac{S_A}{V_A})\cdot r(\frac{C_B}{V_B}+1)+\frac{S_A}{V_A}\cdot(\frac{C_B}{V_B}+1)$$

Because the organic compositions of capital are the same in the two nations, we can substitute C_A/V_A for C_B/V_B to obtain the following result:

$$(\frac{C_A}{V_A}+1+\frac{S_B}{V_B})\cdot r(\frac{C_A}{V_A}+1)+\frac{S_B}{V_B}\cdot(\frac{C_A}{V_A}+1) > (\frac{C_A}{V_A}+1+\frac{S_A}{V_A})\cdot r(\frac{C_A}{V_A}+1)+\frac{S_A}{V_A}\cdot(\frac{C_A}{V_A}+1)$$

With some additional algebraic manipulation, this equation may then be reduced to obtain the final result:

$$\frac{S_B}{V_B} > \frac{S_A}{V_A}$$

Therefore, the rate of surplus value in Mexico must exceed the rate of surplus value in the United States given these assumptions.

Emmanuel's finding that rich nations benefit at the expense of poor nations is interesting because it is achieved without any reference to monopoly power or military intervention.[77] As Shaikh argues, Emmanuel demonstrated that "it is not necessary to abandon the law of competition in order to be able to understand the intrinsic determinants of modern imperialism."[78] The result is obtained by means of an analysis of competitive interaction in the marketplace. It thus provides us with a way of thinking about how voluntary exchange in the international sphere can lead to a situation of worsening inequality between nations. Overall, international trade and the export of capital can be combined to explain how developing nations are exploited in the neocolonial era.[79]

Following the Economic News[80]

According to an article that appeared in *The Wall Street Journal* in July of 2015, Germany posted a record trade surplus during the prior month of May. What is impressive about the trade surplus, according to the article, is that Germany achieved this new record in spite of the ongoing Greek debt crisis and the mounting evidence of a slowdown of the Chinese economy. Furthermore, this increasing trend, which has been powered by German exports, began back in the mid-1990s. Interestingly, the author points out that Germany's export success has come at a cost to its trade partners. That is, the intensely competitive nature of the German export sector has made it more difficult for other Eurozone nations to compete in world markets. Instead, they ended up borrowing heavily to bolster domestic demand, which contributed to their serious problems during the 2008 financial crisis. Greece provides a good example of a nation that struggled as German trade boomed. As the article reports, Greece's trade deficit soared as a percentage of GDP in 2007 while Germany's trade surplus also rose to great heights. Drawing upon the insights of New Trade Theory discussed in this chapter, if Germany's export success reflects its capture of a large number of retainable industries, then it may be operating in an area of trade conflict where its own success comes at the expense of trading partners in a zero-sum trade conflict. On the other hand, if Germany goes far enough in capturing retainable industries from its neighbors, then it may actually enter an area where its own GDP declines with its trading partners. That is, it may undermine its Eurozone neighbors' competitiveness to the point where it actually reduces its own growth prospects. According to the author of the article, Germany's demands that Greece repay the money previously loaned to it might be preventing Germany from restoring more balanced trade relationships. Specifically, by insisting that Greece not cut taxes or increase government spending, domestic aggregate demand cannot be pumped up and so Greece must continue to rely on its exports, which are not as competitive.

Summary of Key Points

1. According to neoclassical economists, even if a nation has an absolute advantage in the production of two commodities, it may still benefit from trade if another country has a comparative advantage.
2. When resources are heterogeneous, a nation experiences increasing marginal opportunity costs, but when resources are homogeneous, a nation experiences constant marginal opportunity costs.
3. In the case of trade between two nations, the limits to the international terms of trade are the domestic terms of trade in each nation.
4. When two nations specialize in the commodities in which they have comparative advantages, each can consume output combinations along a trading possibilities frontier that lies beyond the production possibilities frontier.
5. In the case of increasing marginal opportunity costs, each nation will only pursue partial

specialization because eventually the marginal opportunity costs become equal.

6. The equilibrium terms of trade in the case of increasing marginal opportunity costs is given by the slope of the ray from the origin that passes through the intersection of the two nations' offer curves.
7. The basis of Ian Fletcher's critique of comparative advantage is the existence of eight hidden assumptions that form the basis of the theory.
8. Retainable industries are industries in which economies of scale are so large that the first nation to reach a large production volume captures and retains the industry.
9. New trade theory suggests that free trade leads to a positive sum game of mutual gains at times and to a zero-sum game of bitter conflict at other times.
10. According to neoclassical economists, import tariffs and import quotas lead to losses of social welfare because higher prices reduce the quantity demanded of consumers and increase the production of inefficient domestic producers.
11. According to dependency theorists, a national metropolis appropriates the surpluses of multiple satellites by using its considerable market power.
12. In the theory of unequal exchange, unequal exchanges in the international marketplace allow a high-wage nation to appropriate excess value from a low-wage nation when both nations possess the same organic compositions of capital

and a higher rate of surplus value exists in the low-wage nation.

List of Key Terms

Comparative advantage

Absolute advantage

Increasing marginal opportunity costs

Heterogeneous resources

Homogeneous resources

Domestic terms of trade

International terms of trade

Limits to the terms of trade

Trading possibilities frontier (TPF)

Autarky

Partial specialization

Offer curves

Free trade

Kaldor-Hicks criterion

Retainable industries

Lockout phenomenon

Positive sum game

Zero-sum game

Rational protectionism

Import tariff

Import quota

Dependency theory

Underdevelopment

Center

Periphery

Metropolises

Satellites

Lumpenbourgeoisie

Monopolistic merchant's capital

Modern monopoly capital

Unequal exchange

Value ratio

Production price ratio

Problems for Review

1. Given the information in the table below, determine which nation has an absolute advantage in the production of each commodity and which nation has a

comparative advantage in the production of each commodity.

Problem for Review: Problem 1

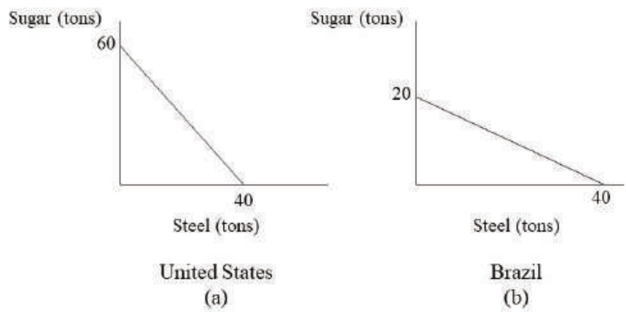
	Sugar (labor hours to produce 1 unit)	Steel (labor hours to produce 1 unit)
U.S.	9	5
Brazil	3	2

2. Given the information in Figure 19.17, answer the following questions:

- Which nation has an absolute advantage in sugar production?
- Which nation has an absolute advantage in steel production?
- Which nation has a comparative advantage in sugar production?
- Which nation has a comparative advantage in steel production?
- What are the limits to the international terms of trade? That is, what is the maximum and minimum price of sugar in terms of steel? What is the maximum and minimum price of steel in terms of sugar?
- If the international terms of trade are 1 unit of sugar/1 unit of steel, then add the new trading

possibilities frontiers to the graphs for each nation. Label the intercepts.

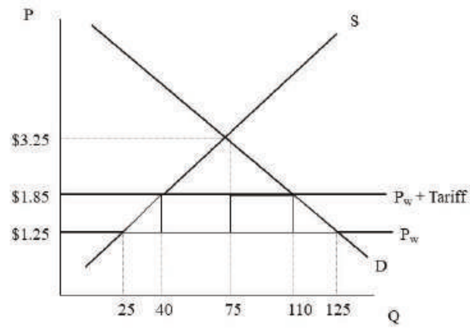
Figure 19.17: Problem #2



3. Given the information in Figure 19.18, answer the following questions:

- If a tariff of $0.60 is imposed as implied in the graph, then calculate the deadweight loss that results from the excess production costs of domestic producers. Hint: Recall the formula for the area of a triangle.
- Calculate the deadweight loss resulting from the reduction in quantity demanded.
- Calculate the tariff revenue.
- If an import quota of 70 units was imposed instead in this case, then which parties might acquire the windfall profit? Why?

Figure 19.18: Problem #3



4. Suppose the two nations represented in the table below produce different commodities and trade in the international marketplace. Complete the table, and calculate the general rate of profit (r). Then calculate the value ratio and the production price ratio. How do these ratios compare? What role do the assumptions of this model play in producing this result?

Nation	Constant Capital	Variable Capital	Surplus Value	Quantity Produced (q)
Nation A	250	275	135	125
Nation B	200	220	75	110
Aggregate				

Notes

1. See Hunt (2002), p. 191-192, for a summary of Mill's theory of international prices.
2. See Fletcher (2009), p. 105-118.
3. I am deeply grateful to Ian Fletcher for granting me permission to include this summary of his perspective. The original source is: Fletcher, Ian. *Free Trade Doesn't Work: Why American Needs a Tariff*. U.S. Business & Industry Council: Washington, DC, 2009. p. 105-118 and 215-231. Information about the most recent edition of his book may be found at: http://www.freetradedoesntwork.com/index.html
4. Ibid. p. 105-106.
5. Ibid. p. 105.
6. Ibid. p. 105.
7. Ibid. p. 105.
8. Ibid. p. 106.
9. Ibid. p. 106-107.
10. Ibid. p. 107-109.
11. I rephrased this hidden assumption. Fletcher (2009), p. 109-110, states it somewhat differently.
12. Ibid. p. 110-113.
13. Ibid. p. 113-115.
14. Ibid. p. 114.
15. Ibid. p. 114.
16. Ibid. p. 115.
17. Noam Chomsky offers a similar critique of Ricardo's original example involving England and Portugal in Chomsky (2002), p. 254.

18. Fletcher (2009), p. 115-118.
19. Ibid. p. 116.
20. Ibid. p. 116.
21. Ibid. p. 215.
22. Ibid. p. 215.
23. Ibid. p. 215.
24. Ibid. p. 216.
25. Ibid. p. 216.
26. Ibid. p. 216.
27. Carbaugh (2011), p. 88, makes the same point using a single, similarly shaped AC curve for two different nations. Although the firms have the same AC curves, the firm that produces the larger volume first ends up with a lower AC.
28. Fletcher (2009), p. 217.
29. Ibid. p. 220.
30. Ibid. p. 221.
31. Ibid. p. 221.
32. Ibid. p. 222.
33. Ibid. p. 222.
34. This graph is a modified version of the one that Fletcher (2009), p. 221, provides. A similar graph can be found in Gomory and Baumol (2000), p. 31.
35. Ibid. p. 224.
36. This graph is a modified version of the one that Fletcher (2009), p. 224, provides. A similar graph can be found in Gomory and Baumol (2000), p. 37.
37. Ibid. p. 223.
38. Ibid. p. 223-225.
39. Ibid. p. 228.

40. Ibid. p. 228.
41. Ibid. p. 229.
42. Ibid. p. 229.
43. Ibid. p. 229.
44. Ibid. p. 229.
45. Ibid. p. 229.
46. Ibid. p. 230-231.
47. Carbaugh (2011), p. 158.
48. Ibid. p. 158.
49. Ibid. p. 158.
50. Rose (2016).
51. Rose (2016) provides a simple diagram to demonstrate these relationships.
52. Frank (1969), p. 3.
53. Howard and King (1992), p. 177.
54. Frank (1969), p. 7.
55. Brewer (1992), p. 164.
56. Ibid. p. 196.
57. Ibid. p. 164.
58. Howard and King (1992), p. 177.
59. Howard and King (1992), p. 177.
60. Frank (1969), p. 7.
61. Brewer (1990), p. 166.
62. Ibid. p. 166.
63. Ibid. p. 166.
64. Ibid. p. 166.
65. Antunes (2017)
66. Antunes (2017)

67. Antunes (2017)
68. Itoh (2009), p. 207.
69. Howard and King (1992), p. 189.
70. Emmanuel (1972), p. 64.
71. Ibid. p. 64.
72. Brewer (1990), p. 209.
73. Brewer (1990), p. 203.
74. Ibid. p. 203.
75. Howard and King (1992), p. 190-191.
76. Howard and King (1992), p. 191.
77. Brewer (1990), p. 200.
78. Shaikh (1980), p. 210.
79. Itoh (2009), p. 207.
80. Mattich, Alen. "Why Germany's Record Surplus Isn't Helping the Eurozone." *The Wall Street Journal*. July 9, 2015.

Chapter 20

BALANCE OF PAYMENTS ACCOUNTING AND THEORIES OF CURRENCY MARKETS

Goals and Objectives:

In this chapter, we will do the following:

1. *Develop* the fundamentals of balance of payments accounting
2. *Explain* how to work with foreign exchange rates
3. *Analyze* the determination of equilibrium exchange rates using supply and demand
4. *Investigate* the Theory of Purchasing Power Parity
5. *Distinguish* between fixed and floating exchange rate regimes
6. *Explore* Marxist theories of imperialist finance and exchange rate determination

This final chapter of the book considers theories of international finance. As the reader may have noticed, not much was said about the role of money in the previous chapter. Curiously, neoclassical economists have a tradition of separating the field of international economics into these two parts. Part of this chapter concentrates on the accounting method, referred to as balance of payments accounting, which is used to describe the financial flows between one nation and the rest of the world. The neoclassical model of supply and demand, as applied to the foreign exchange market or currency market, is also developed in this chapter to explain how exchange rates are determined in competitive markets. Using this framework, we will explore the role of exchange rate adjustments in equalizing purchasing power across nations. We will also consider the role of the central bank in foreign exchange markets and see why currency crises sometimes occur. Finally, we will examine Marxist theories of imperialist finance to provide us with an alternative perspective on the same topic.

Balance of Payments Accounting

When a nation interacts with the rest of the world, it engages in a wide variety of transactions. A nation buys and sells goods and services in international commodity markets. It receives interest payments and pays interest in international financial markets. Domestic employers in that nation also pay foreign employees just as foreign employers sometimes pay domestic employees. Public and private transfer payments are often made from a nation to other nations, or such transfers are received by a nation. Investors in a nation also often buy financial assets like stocks and bonds or the investors sell them to other nations. Economists find it necessary to have a way of recording all these transactions so that they can study how these interactions influence the overall health and direction of a nation's economy.

Balance of payments accounting refers to the method of accounting for all international transactions that occur between one nation and the rest of the world. Its counterpart at the domestic level is **national income accounting**, which we thoroughly analyzed in Chapter 12. The method involves recording each transaction between a nation and the rest of the world as a debit or a credit, depending on the nature of the transaction. For example, a transaction that leads to a payment to the United States from another nation is recorded as a **credit** in the U.S. Balance of Payments and is thus given a positive value. Any transaction that leads to a payment from the United States to another nation is recorded as a **debit** in the U.S. Balance of Payments and is thus given a negative value.

Table 20.1 shows a summary of the U.S. Balance of Payments for 2014.

Table 20.1: The U.S. Balance of Payments for 2014 (in millions of dollars)

Current Account	**Credit/Debit**
Exports of Goods and Services	2,343,205
Primary Income Receipts (Investment Income/Compensation)	823,353
Secondary Income (Current Transfer Receipts)	140,016
Imports of Goods and Services	-2,851,529
Primary Income Payments (Investment Payments/Compensation)	-585,169
Secondary Income (Current Transfer) Payments	-259,202
CURRENT ACCOUNT BALANCE	-389,526
Capital Account	
Capital Transfer Receipts and Other Credits	0
Capital Transfer Payments and Other Debits	-45
CAPITAL ACCOUNT BALANCE	-45
Financial Account	
Net U.S. Acquisition of Financial Assets Excluding Derivatives	-792,145
Net U.S. Incurrence of Liabilities Excluding Derivatives	977,421
Financial Derivatives Other Than Reserves, Net Transactions	54,372
FINANCIAL ACCOUNT BALANCE	239,648
Statistical Discrepancy	149,923
Capital and Financial Account Balance (includes statistical discrepancy)	389,526
U.S. Payments Balance	0

Source: U.S. Bureau of Economic Analysis[1]

It is helpful to think about which types of transactions lead to credits or debits in the U.S. Balance of Payments. For example, when the U.S. exports goods and services, these transactions are recorded as credits because these sales lead to payments to U.S. sellers. When an American investor receives an interest payment from a foreign borrower, the transaction is recorded as a credit as well. When an American worker receives compensation for work performed by a foreign firm, the payment is recorded as a credit. If an American family receives private transfer payments (referred to as remittances) from a family member working in another nation, the transfer payment is recorded as a credit. Each of these transactions is recorded in a subaccount of the Balance of Payments referred to as the **current account**. The current account only includes transactions that do not involve the purchase or sale of an income-earning asset.

Just as credits may be recorded in the current account, debits may also be recorded. For example, when the U.S. imports goods and services, these transactions are recorded as debits in the U.S. current account because these purchases lead to payments from U.S. buyers to foreign sellers. If an American borrower makes an interest payment to a foreign lender, then the payment is recorded as a debit in the current account. If a U.S. employer pays a salary to a foreign worker, then the payment is treated as a debit in the current account. If the U.S. government grants monetary aid to a foreign nation, then the public transfer payment is treated as a debit in the U.S. current account. Once all the credits and debits are added up in the U.S. Current Account, the final calculation is referred to as the **Current Account Balance**. When this figure is positive, it is said that a **current account surplus** exists. When this figure is negative, it is said that a **current account deficit** exists. Table 20.1 shows that a current account deficit of over $389.5 billion existed in 2014.

One subaccount of the Balance of Payments is worth mentioning because it is frequently cited in the financial news. The **trade balance**, as it is commonly known, refers to the difference between exports of goods and services and imports of goods and services. In this case, imports exceed exports of goods and services by about $508 billion. Therefore, a negative U.S. trade balance of approximately $508 billion existed in 2014. When a negative trade balance exists, it is referred to as a **trade deficit**. If exports exceed imports of goods and services, then a positive trade balance exists, which is referred to as a **trade surplus**. If the balance happens to be exactly zero, which would be extremely unlikely in practice, then **balanced trade** exists.

In addition to the current account, another important subaccount is called the **capital account**. Transactions that are recorded in the capital account involve the purchase and sale of nonfinancial assets like copyrights, patents, and trademarks. When these transactions involve sales of nonfinancial assets, they are recorded as credits. When they involve purchases of nonfinancial assets, they are recorded as debits. For example, if a copyright owned by an American firm is sold to a Chinese firm, then the payment received by the American firm is recorded as a credit in the U.S. Capital Account. If a patent owned by a German firm is sold to an American firm, then the payment made by the American firm is recorded as a debit in the U.S. Capital Account.

The final subaccount of interest is the **financial account**. Transactions that are recorded in the financial account involve the purchase and sale of financial assets like stocks and bonds. Examples include purchases and sales of corporate and government bonds, foreign direct investment, and central bank purchases and sales of securities. When these transactions involve sales of financial assets, they are recorded as credits. When they involve purchases of financial assets, they are recorded as debits. For example, if an American bondholder sells a U.S. government bond to a foreign investor, then the payment that the bondholder receives is recorded as a credit in the U.S. Financial Account. If a foreign investor purchases newly issued shares of stock in an American company, then the transaction is recorded as a credit in the financial account because an American firm receives a payment. If the U.S. Federal Reserve sells securities to the Bank of Japan in exchange for Japanese yen, then the transaction is recorded as a credit in the U.S. Financial Account. In the last case, **official reserve holdings** increase. Official reserves include items like foreign currencies and gold held by a nation's central bank.

On the other hand, if an American investor buys a bond from a foreign bondholder, then the payment to the foreign bondholder is recorded as a debit in the U.S. Financial Account. Similarly, if an American investor purchases a controlling stake in a foreign corporation, the payment to the corporation for the shares is recorded as a debit in the U.S. Financial Account. Finally, if the Bank of England sells U.S. Treasury bonds to the U.S. Federal Reserve, then the payment that the Fed makes to the Bank of England is recorded as a debit in the U.S. Financial Account. If the Fed makes the purchase using British pounds, then U.S. official reserves fall by an equal amount. Overall, if the credits in the capital account and financial account exceed the debits in the capital account and financial account, then a **capital and financial account surplus** exists. On the other hand, if the debits in the capital account and financial account exceed the credits in the capital account and financial account, then a **capital and financial account deficit** exists.

We are now in a much stronger position to understand the information presented in Table 20.1. The table shows that a current account deficit existed in the U.S. in 2014. Similarly, the table reveals that a capital and financial account surplus also existed in the U.S. in 2014. In theory, the current account and the capital and financial account should always add up to zero. That is, the overall U.S. Balance of Payments should be balanced with neither a surplus nor a deficit. The reason is that balance of payments accounting uses a double entry bookkeeping method. For every credit that is entered, a corresponding debit must be entered. Similarly, for every debit that is entered, a corresponding credit must be entered. For example, suppose that a U.S. company sells computers to a firm in France. As we have seen, exports are recorded as credits in the balance of payments because a payment is received for the exported goods. What is the offsetting debit in the U.S. Balance of Payments? Suppose that the French company pays for the computers with Euros. The America firm will then deposit the Euros in a Euro-denominated bank account in, let's say, France. This transaction requires a payment to the French bank in exchange for that financial asset (i.e., the bank deposit). Therefore, the transaction is recorded as a debit in the financial account in an amount that exactly offsets the credit associated with the exports in the current account. All transactions that occur between the home nation and the rest of the world can be analyzed in the same manner. As a result, all debits and credits should cancel out in the aggregate, even though each subaccount may show a surplus or deficit.

This result can be proven as follows. Suppose that we add up all the credits in the current account and all the credits in the capital and financial account. Now suppose that we add up all the debits in the current account and all the debits in the capital and financial account. Given what has been said about every credit having a corresponding debit and vice versa, it must be true that the sum of all credits equals the sum of all debits. If CA refers to the current account, KA+FA refers to the capital and financial account, and the subscripts refer to credits (c) and debits (d) in that specific subaccount, then the following must hold:

$$CA_c + (KA + FA)_c = CA_d + (KA + FA)_d$$

Rearranging the terms in the above equation yields the following result:

$$CA_c - CA_d = (KA + FA)_d - (KA + FA)_c$$

In other words, a current account surplus (an excess of credits over debits in the current account) necessarily implies a capital and financial account deficit (an excess of debits over credits in the capital and financial account). Similarly, a current account deficit (debits > credits in the current account) necessarily implies a capital and financial account surplus (credits > debits in the capital and financial account).

Although the balance of payments cannot show a surplus or deficit overall, in practice, statisticians must estimate the total credits and debits in each subaccount. Because statistical estimates do not guarantee completely accurate results, the current account balance is never exactly offset by the capital and financial account balance. Therefore, the difference must be included, which is referred to as **statistical discrepancy**. Once we add this figure in Table 20.1, the surplus in the capital and financial account exactly offsets the current account deficit.

The Net International Investment Position

The balance of payments is a record of all international *flows* of goods and services, interest payments, transfers, and financial and non-financial assets between one nation and the rest of the world. Because the items on the statement are all flows, they are measured during a given period, such as a year or a quarter. Economists are also interested in keeping a record of the *stocks* of all foreign financial assets and foreign financial liabilities that a nation possesses at any one time. The difference between a nation's foreign financial assets and its foreign financial liabilities at a point in time is called its **net investment position**. The statement that contains this information is called the **Net International Investment Position**. It is a bit like a balance sheet for the entire nation, and the net investment position is somewhat like the net worth of the nation.

The U.S. Net International Investment Position for 2014 is shown in Table 20.2.

Table 20.2: The U.S. Net International Investment Position for 2014 (in millions of dollars)

Type of Investment	Amount
U.S. ASSETS	
Direct Investment at Market Value	7,124,034
Portfolio Investment	9,572,539
Financial Derivatives Other Than Reserves	3,224,535
Other Investment	4,240,188
Reserve Assets	434,251
U.S. LIABILITIES	
Direct Investment at Market Value	6,228,795
Portfolio Investment	16,917,146
Financial Derivatives Other Than Reserves	3,150,688
Other Investment	5,318,617
Total U.S. Assets	24,595,547
Total U.S. Liabilities	31,615,246
Net International Investment Position	-7,019,699

Source: U.S. Bureau of Economic Analysis[2]

As Table 20.2 shows, the net international investment position for the U.S. in 2014 was a negative value exceeding $7 trillion. This amount was the net foreign debt of the United States in 2014. This figure means that U.S. liabilities with respect to the rest of the world exceeded U.S. ownership of foreign assets. The U.S. was considered a **net debtor** in 2014 because it owed more to the rest of the world than was owed to it. If its foreign-owned assets exceeded its foreign liabilities, then it would be regarded as a **net creditor**.

An interesting relationship exists between the Balance of Payments and the Net International Investment Position of a nation. Specifically, a nation that receives more from the rest of the world than it spends runs a current account surplus. Its corresponding capital and financial account deficit implies that it lends this amount to the rest of the world by purchasing foreign assets. The increase in its stock of foreign assets increases its net investment position and moves it in the direction of net creditor status.

Similarly, a nation that spends more than it receives from the rest of the world runs a current account deficit. Its corresponding capital and financial account surplus implies that it borrows this amount from the rest of the world by selling assets to foreign investors. The increase in its foreign liabilities worsens its net investment position and moves it in the direction of net debtor status. As a result, the current account deficit that the U.S. experienced in 2014 worsened the U.S. net international investment position by this amount. In other words, the current account deficit in 2014 increased America's net foreign debt.

It is possible to use the national income accounts identity from Chapter 12 to show how a current account balance relates to a nation's government budget and private sector gaps. The identity below shows how a nation's GDP (Y) may be calculated as the sum of the different components of spending on final goods and services.

$$Y = C + I + G + NX$$

In the above identity, C refers to personal consumption expenditures, I refers to gross private domestic investment, G refers to government expenditures, and NX refers to expenditures on net exports (i.e., exports minus imports).

Subtracting taxes (T) and consumer spending (C) from both sides of the equation and substituting X – M for NX yields the following result.

$$Y - T - C = I + G - T + X - M$$

The left-hand side of the equation represents total saving, which is equivalent to disposable income (Y-T)

minus consumer spending. We can thus substitute saving (S) into the equation to obtain the following result.
$$S = I + G - T + X - M$$

Solving for the trade balance (i.e., net exports), we can write the equation as follows:
$$X - M = (S - I) + (T - G)$$

If we ignore primary and secondary income payments and receipts (i.e., investment payments, compensation, and transfers), then the nation's trade balance is the same as the current account balance. Therefore, the current account balance equals the sum of the private sector gap and the government budget gap. For example, if a current account surplus exists, then the nation is a net lender for the year, strengthens its net international investment position, and may achieve this goal with positive net saving in the private sector (S > I) and a government budget surplus (T > G). Net saving in the private and public sectors thus corresponds to net lending in the world economy. On the other hand, if a current account deficit exists, then the nation is a net borrower for the year, worsens its net international investment position, and may bring about this result with negative net saving in the private sector (S < I) and a government budget deficit (T < G). Net borrowing in the private and public sectors thus corresponds to net borrowing in the world economy.

Working with Foreign Exchange Rates

Now that we have an accounting framework for thinking about a nation's international transactions, we will shift gears and begin thinking about how foreign exchange markets work. Once we have a theory that helps us understand how these markets function, we will circle back and connect that theory to balance of payments accounting.

Before we explore the inner workings of foreign exchange markets, we need to consider what foreign exchange rates are and what they mean. On any given day, it is possible to turn to the financial news section of *The Wall Street Journal* and see the current exchange rates. For example, Figure 20.1 provides information about current spot rates (foreign exchange rates) for two consecutive days in January 2008.[3]

Figure 20.1: Foreign Exchange Rates

Currency Trading Exchange Rates (Spot Rates) January 15-16, 2008				
	U.S. $ per unit of foreign exchange ($/FX)		Foreign currency per U.S. $ (FX/$)	
Currency	Tue	Wed	Tue	Wed
Euro	1.4842	1.4643	0.6738	0.6829
Chinese Yuan	0.1382	0.1381	7.2345	7.2400
U.K. Pound	1.9671	1.9638	0.5084	0.5092

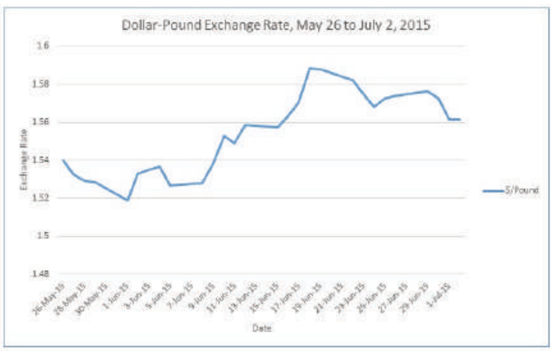
Foreign exchange rates simply tell us what the price of one nation's currency is in terms of another nation's currency. For example, on Tuesday, January 15, 2008, the price of one Euro was about $1.4842. By Wednesday, January 16, 2008, the price of a Euro was only $1.4643. The price of a Euro had thus fallen from one day to the next. The prices of the Chinese Yuan and the U.K. Pound had also fallen. The Yuan fell from $0.1382 to $0.1381, and the U.K. Pound fell from $1.9671 to $1.9638. All these currencies fell in value relative to the U.S. dollar over the course of these two days. When a currency's value declines relative to another currency, we say that it has **depreciated** against the other currency.

If we consider the two columns on the right in Figure 20.1, we see a rather different result. From Tuesday to Wednesday, each of these numbers increased. For example, the price of $1 on Tuesday, January 15, 2008 was 0.6738 Euros and on Wednesday it was 0.6829 Euros. That is, the dollar increased in value relative to the Euro. The situation was similar with the other two currencies. The U.S. dollar increased in value from 7.2345 to 7.2400 Yuan and from 0.5084 to 0.5092 pounds. Whenever a currency increases in value relative to another, we say that the currency has **appreciated** relative to the other currency.

It is no coincidence that each of these foreign currencies depreciated against the U.S. dollar while the U.S. dollar appreciated against them. In fact, it must be the case. The reader may have noticed that the calculations in the two columns on the left are in terms of U.S. dollars per unit of foreign exchange ($/FX) whereas the calculations in the two columns on the right are in terms of foreign exchange per U.S. dollar (FX/$). The one calculation is the reciprocal of the other. Therefore, if an increase occurs from one day to the next in the first two columns, then a decrease must occur from one day to the next in the final two columns, and vice versa.

For these exchange rates to change from one day to the next is not exceptional. In fact, foreign exchange rates typically change from minute to minute. Consider, for example, the U.S. dollar-U.K. pound exchange rate during the period from May 26 to July 2, 2015 shown in Figure 20.2.

Figure 20.2: Fluctuations in the Dollar-Pound Exchange Rate



Source: Board of Governors of the Federal Reserve System. See Economics Research & Data, Statistical Releases and Historical Data, Foreign Exchange Rates, Country Data at: www.federalreserve.gov

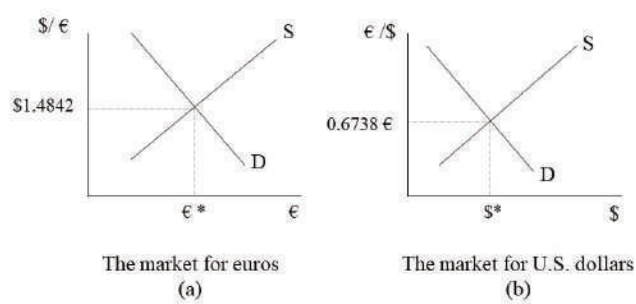
The pound fluctuated during that short period from less than \$1.52 to almost \$1.59. Much larger fluctuations can sometimes occur, as we will discuss later in the chapter.

Explaining Foreign Exchange Rate Fluctuations

It is natural to wonder why these fluctuations of exchange rates occur. The answer is that exchange rates are determined competitively in international currency markets or foreign exchange markets. As buyers and sellers change their purchases and sales for a huge variety of reasons, the market prices of foreign currencies change in response. To explain market prices in Chapter 3, we used the supply and demand model. In this chapter, we will once again turn to the supply and demand model to understand how **equilibrium exchange rates** and **equilibrium quantities exchanged** are determined.

In the market for any currency, we can speak about the supply and demand for that currency. The sellers of the currency represent the supply side of the market, and the buyers of the currency represent the demand side of the market. Their competitive interaction determines the exchange rate and the amount exchanged. Figure 20.3 (a) offers an example of supply and demand in the market for Euros.

Figure 20.3: Two Perspectives of Foreign Exchange Markets



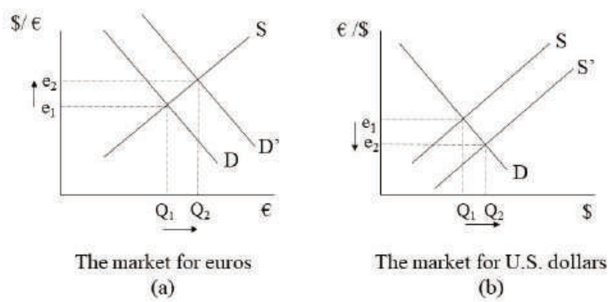
The equilibrium exchange rate of $1.4842 per Euro occurs where the quantity demanded and the quantity supplied of Euros are equal. If the exchange rate exceeds this level, then a surplus of foreign exchange will exist, and competition will drive the exchange rate down. If the exchange rate falls short of this level, then a shortage of foreign exchange will exist, and competition will drive the exchange rate up. The equilibrium quantity exchanged of Euros (€*) is also determined as competition brings about the equilibrium outcome in the market.

It is possible to present a second perspective of the foreign exchange market. Let's consider the market for U.S. dollars. Figure 20.3 (b) offers an example of supply and demand in the market for dollars. The equilibrium exchange rate of 0.6738 Euros per dollar occurs where the quantity demanded equals the quantity supplied of dollars. As before, surpluses and shortages are eliminated through competition and bring about an equilibrium Euro-U.S. dollar exchange rate and an equilibrium quantity exchanged of dollars ($*). The reader should notice that these two exchange rates are reciprocals of each other. In fact, they represent the same market viewed from two different perspectives. Therefore, if the dollar-euro market is in equilibrium, then it must be the case that the euro-dollar market is also in equilibrium, as shown in Figure 20.3.

Also of importance is the relationship between the supply and demand curves in each market. The sellers supplying Euros in the Euro market must also be demanding U.S. dollars in the dollar market. Similarly, the buyers of Euros in the Euro market must also be supplying U.S. dollars in the dollar market. Think about an American who goes to Europe on vacation. She wants to buy some souvenirs but needs Euros to do so. She buys Euros and must sell U.S. dollars to do so. Similarly, a European tourist in the United States must sell Euros to buy U.S. dollars. To sell one is to buy the other and to buy one is to sell the other. In summary, the two markets are mirror reflections of one another.

The mirror reflection argument applied to foreign exchange markets is consistent with what we said earlier about the reciprocal relationship between the exchange rates. For example, suppose that the demand for Euros increases, shifting the demand curve for Euros to the right in Figure 20.4 (a).

Figure 20.4: The Reciprocal Interaction Between Exchange Rates



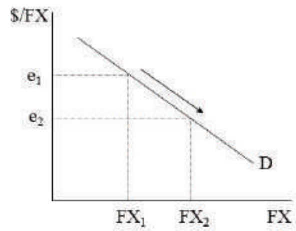
The rightward shift of demand in the market for Euros corresponds to a rightward shift of supply in the market for U.S. dollars. As mentioned previously, buyers of Euros are sellers of U.S. dollars when we are discussing the dollar-euro and euro-dollar markets. The consequence is an appreciation of the Euro and a depreciation of the U.S. dollar, reflecting the reciprocal connection between these exchange rates. The equilibrium quantities exchanged of both Euros and U.S. dollars increase as well.

Building the Supply and Demand Model of Foreign Exchange Markets

It is easy to declare that supplies and demands for foreign currencies exist, but we really need to build the supply and demand model of foreign exchange markets from scratch to understand the underlying determinants of exchange rate movements. We will begin with the demand side of a foreign currency market. As stated previously, we can consider the market for foreign exchange or the market for the domestic currency (e.g., U.S. dollars) since each is just a mirror reflection of the other. Throughout the remainder of this chapter, we will adopt the perspective of a market for foreign exchange. The reason for doing so is that we are accustomed to thinking of products and services as having dollar prices. For example, the price of an automobile is stated in terms of dollars per automobile ($/auto). We can think of the price of Japanese Yen in the same way. That is, the price of the Yen (¥) is in terms of so many dollars per Yen ($/¥). If the exchange rate (e) rises, then the Yen appreciates, just as a rise in the price of an automobile would mean that it has appreciated in value. Whenever we consider exchange rates, therefore, we will refer to dollars per unit of foreign exchange ($/FX).

Figure 20.5 shows a downward sloping demand curve for foreign exchange.

Figure 20.5: The Quantity Demanded of Foreign Exchange



As the exchange rate falls, the foreign currency becomes cheaper and the quantity demanded of foreign exchange rises. It is worth asking why the demand curve slopes downward in this case.[4] When we learned about product markets, it seemed obvious that a lower price for a good, like apples, would lead to a rise in the quantity demanded of apples, ceteris paribus. As the price falls, other things the same, consumers are willing and able to buy more apples. The problem is that the explanation does not seem quite so obvious in the case of a foreign currency. After all, we do not eat Euros and so a lower price may not seem like it would automatically lead to increased purchases. Although we do not eat foreign currencies, they are useful as a means to an end. That is, if we wish to buy foreign goods and services, then we need to buy foreign currencies. Therefore, when the price of a foreign currency declines, we buy more of it because it allows us to purchase more foreign goods and services. That is, foreign goods and services become cheaper when the exchange rate ($/FX) falls and so we purchase more of the foreign currency.

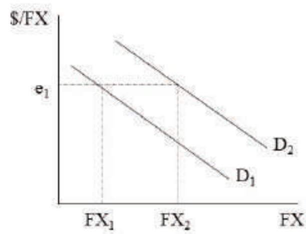
In addition to imported goods and services becoming cheaper when the exchange rate falls, it is also true that foreign assets become cheaper when the exchange rate falls. That is, foreign stocks, bonds, and businesses become less expensive and so investors buy more foreign currency when its price falls. In this case, foreign currency is viewed as a means to buy foreign assets.

A final reason why the demand curve for foreign exchange slopes downward deals with currency speculation. That is, when the price of a foreign currency falls, speculators may expect it to rise later to some normal level. Therefore, speculators will purchase more as the exchange rate falls because they will anticipate a **capital gain** from the later appreciation of the currency.[5]

These three factors combine to create a downward sloping demand for foreign exchange. We could easily reverse the logic for all three factors to explain what happens when the exchange rate rises. When the exchange rate rises, the foreign currency becomes more expensive. Therefore, imports of goods and services become more expensive as do foreign assets. Furthermore, speculators will expect the exchange rate to fall later and so they expect a **capital loss** in the future. As a result, speculators will buy less foreign exchange.

We have now explained movements along the downward sloping demand curve for foreign exchange, but shifts of the demand curve for foreign exchange are possible too. Figure 20.6 shows a rightward shift of the demand curve for foreign exchange.

Figure 20.6: Changes in the Demand for Foreign Exchange



In Figure 20.6 a rise in the quantity demanded of foreign exchange occurs at every possible exchange rate. Which factors might cause such a shift of the demand curve for foreign exchange?[6] The first factor that we will consider is a change in *consumers' preferences* in the home nation. For example, let's suppose that American consumers develop a stronger preference for a specific nation's automobiles. In that case, the demand for that foreign currency will increase. That is, the quantity demanded of the foreign currency will rise at every exchange rate, resulting in a rightward shift of the demand curve. If for some reason, the preferences of Americans changed such that they wanted fewer automobiles at each exchange rate, then the demand curve would shift to the left.

A second factor that might shift the demand curve for foreign exchange is a change in *consumers' incomes* in the home nation. For example, suppose that Americans experience a rise in their incomes due to an economic expansion. They will demand more goods and services as well as more assets. This greater demand for goods, services, and assets, includes a greater demand for the goods, services, and assets of foreign nations. In this case, we would expect a higher demand for foreign exchange and a rightward shift of the demand curve for foreign exchange. On the other hand, if Americans experience a drop in their incomes due to a recession, then they will demand fewer goods, services, and assets, including the goods, services, and assets of other nations. As a result, the demand for foreign exchange will decline, and the demand curve for foreign exchange will shift to the left.

A third factor that might shift the demand curve for foreign exchange is a change in the *relative price levels* of the two nations. For example, suppose that the price level in the U.S. rises while the price level in the foreign nation remains the same. In that case, Americans will view foreign goods, services, and assets as relatively cheaper. Therefore, they will demand a great amount of foreign exchange to buy these relatively less expensive foreign commodities and assets. The same result would occur if the U.S. price level remained the same and the foreign price level fell. On the other hand, a drop in the U.S. price level and/or a rise in the foreign price level would lead to a reduction in the demand for foreign exchange because U.S. goods, services, and assets would now be relatively cheaper than foreign commodities and assets.

A fourth factor that might shift the demand curve for foreign exchange is a change in the *relative interest rates* of the two nations. For example, suppose that interest rates in the foreign nation rise while interest rates in the U.S. remain the same. American investors will view foreign assets as better investments because they now pay more interest. The reader should also recall that as foreign interest rates rise, foreign asset prices will fall. Therefore, foreign assets will be viewed as a bargain and the demand for foreign exchange will increase. As a result, the demand curve for foreign exchange will shift to the right. The same result would occur if U.S. interest rates fell while foreign interest rates remained the same. American investors would view foreign interest-bearing assets as the better investments and would thus demand more foreign exchange. On the other hand, if U.S. interest rates rose as foreign interest rates remained the same or fell, then the demand for foreign exchange would fall and the demand curve for foreign exchange would shift to the left as American investors shied away from foreign investments.

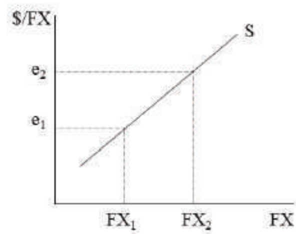
A fifth and final reason that we will consider for shifts of the demand curve for foreign exchange deals with *currency speculation*. For example, suppose that the expectations of speculators change such that the exchange rate is expected to rise soon. In other words, they expect the foreign currency to appreciate soon. In that case, the demand for foreign exchange will rise as speculators decide to buy the currency before it appreciates. As a result, the demand curve for foreign exchange will shift to the right. Alternatively, suppose that speculators decide that the foreign currency is expected to depreciate soon. In that case, speculators will demand less foreign exchange, and the demand curve for foreign exchange will shift to the left. That is, speculators do not want to buy the foreign exchange because they expect it to lose value soon.

In these examples, it is important to keep in mind the distinction between a movement along the demand curve for foreign exchange and a shift of the demand curve. Movements along the demand curve for foreign exchange are caused by changes in the exchange rate. Shifts of the demand curve for foreign exchange, on the other hand, result from changes in any other factors that might affect the buyers of foreign exchange. This distinction is the same one discussed in Chapter 3 between a **change in quantity demanded** and a **change in demand**. Understanding the difference helps prevent possible confusions that might otherwise arise. For example, the reader may have noticed that speculative behavior has been used to explain both a movement along the demand curve and a shift of the demand curve for foreign exchange. In the case of a movement along the demand curve, the speculative response is a direct response to a falling exchange rate. In the case of a shift of the demand curve, speculators experience a change in their expectations that is not related to a change in the current exchange rate.

The situation is very similar on the supply side of the foreign exchange market. Indeed, the same factors should be at play on the supply side since the supply of foreign exchange is the mirror reflection of the demand for dollars. Because the demand for dollars is affected by the factors that we just described, the supply of foreign exchange must be affected by these same factors.

Figure 20.7 shows an upward sloping supply curve for foreign exchange.

Figure 20.7: The Quantity Supplied of Foreign Exchange



As the exchange rate rises, the foreign currency becomes more expensive and the quantity supplied of foreign exchange rises. It is worth asking why the supply curve slopes upward in this case. When we learned about product markets, it seemed obvious that a higher price for a good, like apples, would lead to a rise in the quantity supplied of apples, ceteris paribus. As the price rises, other things the same, producers are willing and able to sell more apples. In the case of foreign exchange markets, sellers are willing and able to sell more foreign exchange because when they sell foreign exchange, they acquire U.S. dollars, which are useful as a means to an end. That is, if foreigners wish to buy American goods and services, then they need to buy U.S. dollars. Therefore, when the price of a foreign currency rises, they sell more of it because it allows them to purchase more American goods and services. That is, American goods and services become cheaper when the exchange rate ($/FX) rises and so they sell more of the foreign currency.

In addition to U.S. exports of goods and services becoming cheaper to foreign consumers when the exchange rate rises, it is also true that American assets become cheaper to foreign investors when the exchange rate rises. That is, American stocks, bonds, and businesses become less expensive and so foreign investors sell more foreign currency when its price rises. In this case, U.S. dollars are viewed as a means to buy American assets.

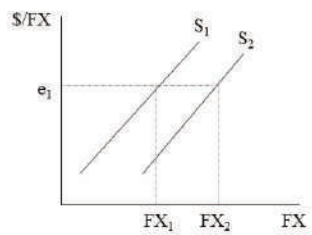
A final reason why the supply curve for foreign exchange slopes upward deals with currency speculation. That is, when the price of a foreign currency rises, speculators may expect it to fall later to some normal level. Therefore, speculators will sell more of the foreign currency as the exchange rate rises because they will anticipate a *capital loss* from the later depreciation of the currency if they do not sell now.

These three factors combine to create an upward sloping supply for foreign exchange. We could easily reverse the logic for all three factors to explain what happens when the exchange rate falls. When the exchange rate falls, the foreign currency becomes less expensive. Therefore, U.S. exports of goods and services become more expensive to foreigners as do foreign assets, and so less foreign exchange will be sold. Furthermore, speculators will expect the exchange rate to rise later and so they expect a *capital gain* in the future. As a result, speculators will decide to sell less foreign exchange.

We have now explained movements along the upward sloping supply curve for foreign exchange, but shifts of the supply curve for foreign exchange are possible too.

Figure 20.8 shows a rightward shift of the supply curve for foreign exchange.

Figure 20.8: Changes in the Supply of Foreign Exchange



In Figure 20.8 a rise in the quantity supplied of foreign exchange occurs at every possible exchange rate. Which factors might cause such a shift of the supply curve for foreign exchange? Just as in the case of the demand for foreign exchange, it is possible that *consumer preferences* in the foreign nation change. For example, let's suppose that foreign consumers develop a stronger preference for American automobiles. In that case, the supply of the foreign currency will increase. That is, the quantity supplied of the foreign currency will rise at every exchange rate, resulting in a rightward shift of the supply curve. If for some reason, the preferences of foreigners changed such that they wanted fewer American automobiles at each exchange rate, then the supply curve would shift to the left.

A second factor that might shift the supply curve for foreign exchange is a change in *consumers' incomes* in the foreign nation. For example, suppose that foreigners experience a rise in their incomes due to an economic expansion. They will demand more goods and services as well as more assets. This greater demand for goods, services, and assets, includes a greater demand for the goods, services, and assets of the United States. In this case, we would expect a higher supply of foreign exchange and a rightward shift of the supply curve for foreign exchange. On the other hand, if foreigners experience a drop in their incomes due to a recession, then they will demand fewer goods, services, and assets, including the goods, services, and assets of the United States. As a result, the supply of foreign exchange will decline and the supply curve for foreign exchange will shift to the left.

A third factor that might shift the supply curve for foreign exchange is a change in the *relative price levels* of the two nations. For example, suppose that the price level in the foreign nation rises while the price level in the U.S. remains the same. In that case, foreigners will view U.S. goods, services, and assets as relatively cheaper. Therefore, they will sell a greater amount of foreign exchange to obtain dollars with which they can buy these relatively less expensive American commodities and assets. The same result would occur if the foreign price level remained the same and the American price level fell. On the other hand, a drop in the foreign price level and/or a rise in the American price level would lead to a reduction in the supply of foreign exchange because foreign goods, services, and assets would now be relatively cheaper than American commodities and assets.

A fourth factor that might shift the supply curve for foreign exchange is a change in the *relative interest rates* of the two nations. For example, suppose that interest rates in the United States rise while interest rates in the foreign nation remain the same. Foreign investors will view U.S. assets as better investments because they now pay more interest. Furthermore, if U.S. interest rates rise, U.S. asset prices will fall. Therefore, U.S. assets will be viewed as a bargain and the supply of foreign exchange will increase. As a result, the supply curve for foreign exchange will shift to the right. The same result would occur if foreign interest rates fell while U.S. interest rates remained the same. Foreign investors would view American interest-bearing assets as the better investments and would thus supply more foreign exchange. On the other hand, if foreign interest rates rose as U.S. interest rates remained the same or fell, then the supply of foreign exchange would fall and the supply curve for foreign exchange would shift to the left as foreign investors shied away from American investments.

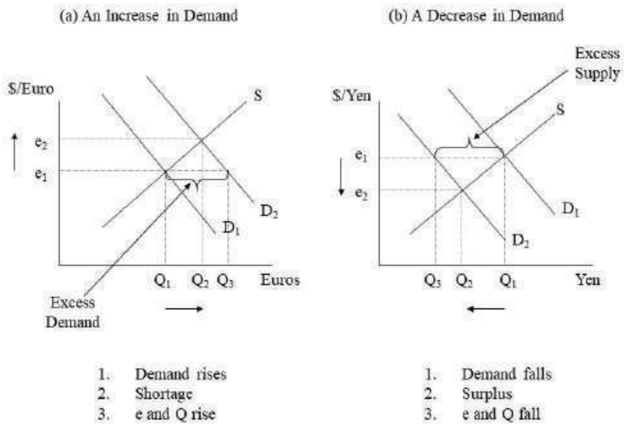
A fifth and final reason that we will consider for shifts of the supply curve for foreign exchange deals with *currency speculation*. For example, suppose that the expectations of speculators change such that the exchange rate is expected to fall soon. In other words, they expect the foreign currency to depreciate soon. In that case, the supply of foreign exchange will rise as speculators decide to sell the currency and avoid potential losses from its future depreciation. As a result, the supply curve for foreign exchange will shift to the right. Alternatively, suppose that speculators decide that the foreign currency is expected to appreciate soon. In that case, speculators will supply less foreign exchange, and the supply curve for foreign exchange will shift to the left. That is, speculators will not want to sell the foreign currency because they expect it to gain value soon.

Once again, the distinction between a movement along the supply curve and a shift of the supply curve is important to keep in mind. Movements along the supply curve are always caused by changes in the exchange rate. Shifts of the supply curve, on the other hand, are always caused by changes in any other factors that influence sellers of foreign exchange aside from exchange rate changes. This distinction is the same one discussed in Chapter 3 between a **change in quantity supplied** and a **change in supply**.

We are now able to combine the supply and demand sides of the foreign exchange market to provide a more thorough explanation of the equilibrium exchange rate and the equilibrium quantity exchanged of foreign exchange. Furthermore, we use the model to explain how exchange rates change in response to exogenous shocks. That is, using the method of comparative statics, we can analyze how changes in consumers' preferences, consumers' incomes, relative price levels, relative interest rates, and the expectations of speculators can cause the equilibrium exchange rate and the equilibrium quantity exchanged to change in a specific foreign exchange market.

For example, suppose that American consumers learn about a new European automobile that is spacious, comfortable, and environmentally friendly. The popularity of this product causes American consumers to increase their demand for Euros so that they can purchase these automobiles. The demand for Euros will shift to the right as shown in Figure 20.9 (a).

Figure 20.9: Cases Involving Changes in Demand

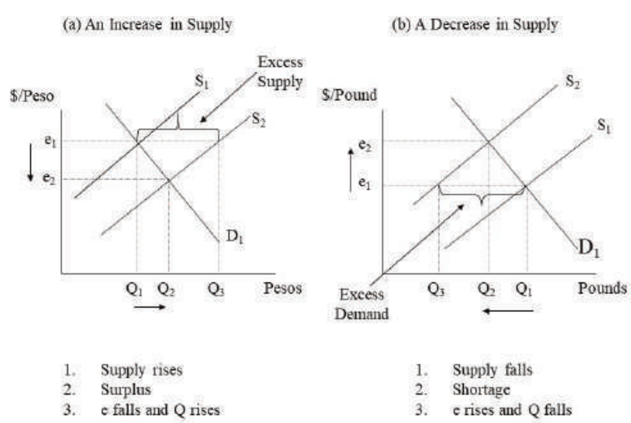


The higher demand for Euros creates a shortage or an excess demand for Euros. As a result, competition drives up the exchange rate and the quantity exchanged of the Euro to their new equilibrium values of e_2 and Q_2. In this case, the Euro appreciates relative to the dollar.

Consider a different example. Suppose that interest rates in Japan fall significantly relative to interest rates in the United States. American investors will find Japanese assets to be less attractive investments, and they will want to purchase fewer of them. As a result, they will demand fewer Yen. The demand for Yen will shift to the left as shown in Figure 20.9 (b). The lower demand for Yen creates a surplus or an excess supply of Yen. Competition then drives the exchange rate and the quantity exchanged of Yen down to their new equilibrium values of e_2 and Q_2. In this case, the Yen depreciates relative to the dollar.

As another example, consider what will happen if speculators conclude that the Mexican peso will depreciate soon. They will want to sell pesos quickly before they lose value, and the supply of pesos will rise in the foreign exchange market. As a result, the supply curve shifts to the right, as shown in Figure 20.10 (a).

Figure 20.10: Cases Involving Changes in Supply



The resulting surplus or excess supply of pesos puts downward pressure on the peso. The exchange rate falls and the quantity exchanged of pesos rises to their new equilibrium values of e_2 and Q_2. In this case, the peso has depreciated. The belief that the peso will depreciate turns out to be accurate, but it is also a **self-fulfilling prophecy**. That is, speculators' belief that the peso will soon depreciate brings about its depreciation. Such self-fulfilling prophecies are not unusual in foreign currency markets and can often create considerable volatility.

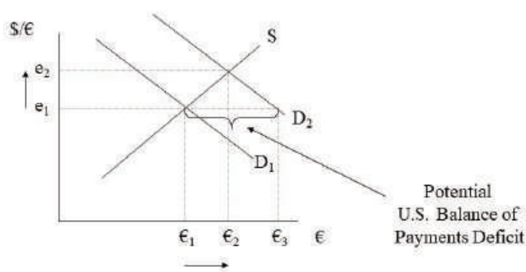
Finally, let's consider a case where the price level in the United States rises relative to the price level in Britain. In this case, British consumers will find American exports of goods and services to be more expensive. As a result, they will demand fewer American exports and so they will supply fewer pounds in the foreign exchange market, as shown in Figure 20.10 (b). The reduced supply of pounds will create a shortage of pounds. Competition for the limited supply of pounds will then drive up the exchange rate and cause the quantity exchanged of pounds to fall overall until they reach their new equilibrium values of e_2 and Q_2. In this case, the pound appreciates relative to the dollar.

It is important to keep in mind that each of these examples is relatively simple. Frequently, many different factors will be changing at the same time, leading to simultaneous shifts in both the supply and demand sides of the market. In fact, even if only one factor changes, both sides of the market will be affected because each supply curve is the mirror reflection of a demand curve. In other words, the same factors affect both sides of the market. Complex cases make it difficult to draw conclusions about the overall impact on the equilibrium outcome, as we learned in Chapter 3. As we discussed in the case of product markets, when both sides of the market are affected, unless we know the extent of the shifts, the direction of the overall change in either the exchange rate or the quantity exchanged will be indeterminate. The reader might want to review the examples of complex cases in Chapter 3 when thinking about how simultaneous shifts of supply and demand might influence the foreign exchange market.

Floating Versus Fixed Exchange Rate Regimes

Now that we have a solid understanding of the supply and demand model of the foreign exchange market, it will be easier to see how this analysis connects to balance of payments accounting. Suppose, for example, that the demand for Euros rises in the U.S. dollar-Euro market, as shown in Figure 20.11.

Figure 20.11: The Elimination of a U.S. Balance of Payments Deficit in the Case of Floating Exchange Rates

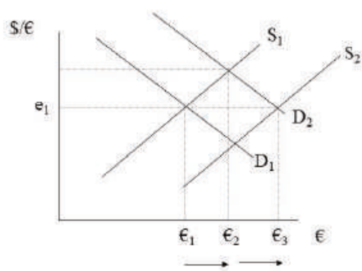


The increase in the demand for Euros creates a shortage of Euros, but we can also give another interpretation to the shortage. The quantity demanded of Euros (€3) at the initial exchange rate of e1 exists because Americans wish to buy European imports of goods and services, European assets, and Euros for speculation. These demands are potential debits in the U.S. Balance of Payments. At the same time, the quantity supplied of Euros (€$_1$) at the initial exchange rate of e_1 exists because Europeans wish to purchase American exports of goods and services, American assets, and U.S. dollars for speculation. These supplies are potential credits in the U.S. Balance of Payments. Therefore, the difference between the quantity demanded of Euros and the quantity supplied of Euros may be interpreted as the potential excess of debits over credits in the U.S. Balance of Payments. That is, it reflects a potential balance of payments deficit. Of course, the overall balance of payments must be balanced. Therefore, we need an explanation for how this discrepancy is resolved.

To understand how the overall balance in the balance of payments is achieved, it is helpful to consider two different types of **exchange rate regime**. That is, we will consider two methods that governments might adopt in their approach to foreign exchange markets. The first type is a **floating exchange rate regime**. In this case, the exchange rate moves to its equilibrium level. For a currency that is widely traded, like the U.S. dollar or Japanese Yen, competition will lead to quick changes in exchange rates whenever the equilibrium level is disturbed by an external shock that affects the supply or demand of the currency. Therefore, when a balance of payments deficit exists, the exchange rate will quickly rise to eliminate the payments deficit and ensure an overall balance in the balance of payments.[7]

Another possible exchange rate regime is referred to as a **fixed exchange rate regime**. When governments and central banks adopt a policy of fixed exchange rates, they commit themselves to a specific value of the exchange rate. If market competition drives the exchange rate up or down, the central bank intervenes by selling or buying the foreign currency to ensure that the exchange rate returns to its original level. This policy affects the central bank's reserve assets. In the case of the potential balance of payments deficit that is represented in Figure 20.11, the shortage will put upward pressure on the Euro. Because the U.S. Federal Reserve is committed to a fixed exchange rate at e_1, however, it will sell Euros in the foreign exchange market. This increase in the supply of Euros will shift the supply curve of Euros to the right, as shown in Figure 20.12.

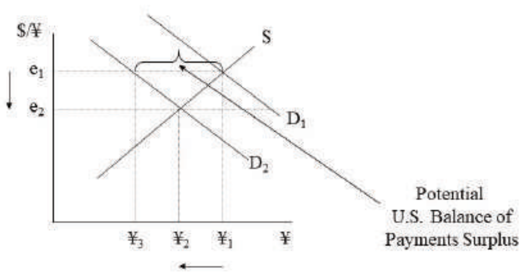
Figure 20.12: A Fixed Exchange Rate Policy in the Case of a Shortage of Euros



This increase in supply will allow the central bank to maintain an exchange rate of e_1. It will also eliminate the potential for a balance of payments deficit because credits and debits will be equal. The difference has been made up with the expenditure of official reserves. That is, Euros held in reserve with the central bank have been spent in the foreign exchange market. These sales of Euros and purchases of U.S. dollars count as credits in the U.S. Balance of Payments. The result is an overall payments balance and a fixed exchange rate. In other words, since exports of U.S. goods, services, and assets are not sufficient to purchase imports of goods, services, and assets, then the difference must be made up by drawing down official reserves.

Next suppose that the demand for Yen falls in the U.S. dollar-Yen market, as shown in Figure 20.13.

Figure 20.13: The Elimination of a U.S. Balance of Payments Surplus in the Case of Floating Exchange Rates



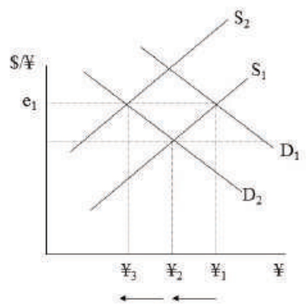
The decrease in the demand for Yen creates a surplus of Yen, but we can also give another interpretation to the surplus. The quantity supplied of Yen (¥1) at the initial exchange rate of e1 exists because Japanese consumers and investors wish to buy U.S. exports of goods and services, U.S. assets, and U.S. dollars for speculation. These supplies are potential credits in the U.S. Balance of Payments. At the same time, the quantity demanded of

Yen (¥3) at the initial exchange rate of e_1 exists because Americans wish to purchase Japanese imports of goods and services, Japanese assets, and Yen for speculation. These demands are potential debits in the U.S. Balance of Payments. Therefore, the difference between the quantity supplied of Yen and the quantity demanded of Yen may be interpreted as the potential excess of credits over debits in the U.S. Balance of Payments. That is, it reflects a potential balance of payments surplus. Of course, the overall balance of payments must be balanced. Therefore, just as before, we need an explanation for how this discrepancy is resolved.

As before, in the case of a floating exchange rate regime, when a balance of payments surplus exists, the exchange rate will quickly fall to eliminate the payments surplus and ensure an overall balance in the balance of payments, as shown in Figure 20.13.

In the case of the potential balance of payments surplus that is represented in Figure 20.13, the surplus will put downward pressure on the Yen. If the U.S. Federal Reserve is committed to a fixed exchange rate at e_1, however, it will reduce its supply of Yen in the foreign exchange market. This reduction in the supply of Yen will shift the supply curve of Yen to the left, as shown in Figure 20.14.

Figure 20.14: A Fixed Exchange Rate Policy in the Case of a Surplus of Yen



This decrease in supply will allow the central bank to maintain an exchange rate of e_1. It will also eliminate the potential for a balance of payments surplus because credits and debits will be equal. The difference has been made up with an increase in U.S. official reserves. That is, Yen have been withdrawn from the foreign exchange market and added to official reserve assets. Because the supply of Yen represents potential credits in the U.S. Balance of Payments, when they are withdrawn, potential credits fall until they match debits. The result is an overall payments balance and a fixed exchange rate. In other words, since exports of U.S. goods, services, and assets exceed U.S. imports of Japanese goods, services, and assets, the surplus of Yen must be added to foreign exchange reserves if the exchange rate is to be kept at the same level.

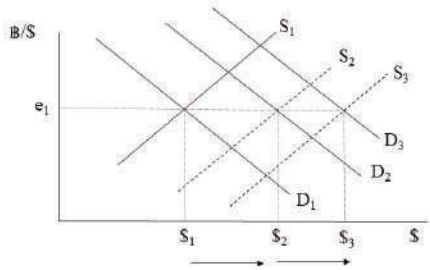
A system of fixed exchange rates was the norm in the years after World War II until the early 1970s. When a government or central bank with a fixed exchange rate decides to reduce its value, it is referred to as a **devaluation** of the currency. When a government or central bank decides to increase its value, it is referred to as a **revaluation** of the currency. Since the early 1970s, a system of **managed floating exchange rates** has become the norm for many nations. This kind of a system allows exchange rates to find their equilibrium levels, but central banks will sometimes intervene to stabilize their currencies. Other nations have simply adopted a floating exchange rate regime.

Currency Crises

In some situations, investors might rapidly sell a currency causing its foreign exchange value to plummet. This situation is referred to as a **currency crisis**. Efforts to stabilize the exchange rate may fail if chronic balance of payments deficits lead to the rapid depletion of foreign exchange reserves. Eventually the central bank may lose control of the exchange rate and then the foreign exchange value of the currency will continue its downward movement. In his book *Globalization and Its Discontents* (2002), Nobel Prize-winning economist Joe Stiglitz explains that the Thai baht collapsed on July 2, 1997 when its value fell overnight by 25% against the U.S. dollar. This event was the trigger for the East Asian financial crisis, which spread to several other nations with effects felt around the world.[8]

Let's analyze this situation by considering the Thai baht-U.S. dollar foreign exchange market, shown in Figure 20.15.

Figure 20.15: Balance of Payments Deficits in Thailand

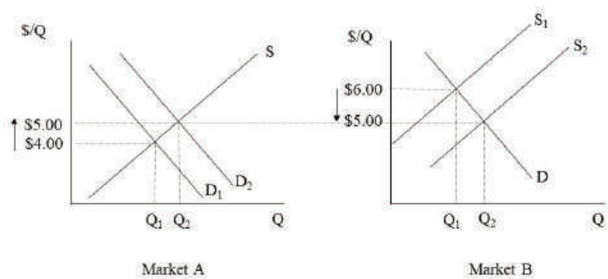


In this case, the demand for the U.S. dollar rises, which creates a shortage of U.S. dollars. The result is a balance of payments deficit in Thailand. If Thailand wishes to prevent a drop in the value of the baht then its central bank must sell dollars, which pushes the value of the dollar back down and the baht back up. If additional increases in the demand for U.S. dollars occur, then the central bank in Thailand must sell even more U.S. dollars to prevent the fall of the baht. Eventually, it will run short of U.S. dollar reserves and may be forced to abandon the peg to the dollar, which is what occurred during the 1997 currency crisis in Thailand.[9]

The Theory of Purchasing Power Parity

In this section, we will consider an additional aspect of neoclassical theory that is referred to as the theory of purchasing power parity. According to the **theory of purchasing power parity**, a single price for a commodity or asset will emerge in a global market economy even though many different currencies exist. To understand how it works, let's first consider a simple case of two regions in the United States where a commodity is bought and sold. Assume that the equilibrium prices in the two regional markets differ. In Figure 20.16, for example, the initial equilibrium price in market A is $4 per unit, and the initial equilibrium price in market B is $6 per unit.

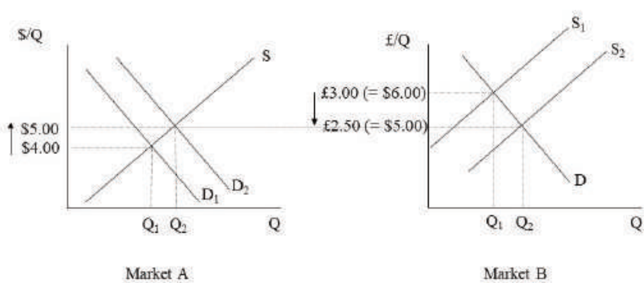
Figure 20.16: A Case of Inter-Regional Arbitrage with a Common Currency



If we ignore transportation costs, it will not take long before some motivated individuals decide to purchase the commodity in market A and then sell the commodity in market B. The higher demand in market A will push up the equilibrium price in market A to $5 per unit, just as the higher supply in market B will push down the equilibrium price in market B to $5 per unit. Eventually, the prices will equalize, as shown in Figure 20.16. This practice of buying a commodity at a low price in one location and selling it at a higher price in another location is called **arbitrage**.

According to the theory of purchasing power parity, arbitrage occurs at the international level as well. Now let's consider an example in which commodities are priced differently in two different markets, but this time each market uses a different medium of exchange. That is, we assume that different currencies are used in the two markets. For example, suppose that the commodity is priced lower in the U.S. and priced higher in the U.K, as shown in Figure 20.17.

Figure 20.17: A Case of International Arbitrage (assuming a fixed exchange rate of \$2/£)



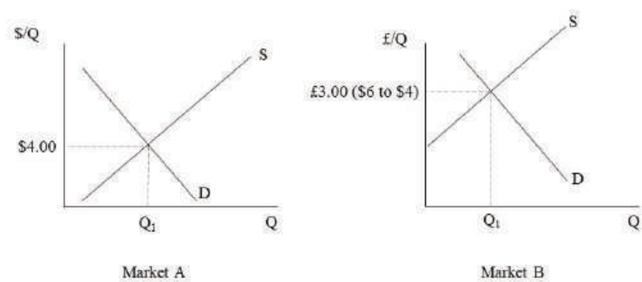
Because of the different currencies, it is necessary to convert one currency into the other so that we can compare prices. This conversion can only be carried out using the current exchange rate. Let's assume that the pound-dollar exchange rate is £0.50/\$ and that the dollar-pound exchange rate is \$2/£. The reader should recall that each is simply the reciprocal of the other. Therefore, if the commodity is priced at £3 in the U.K., then its dollar price must be \$6 (= £3 times \$2/£).

Because the price of the commodity in the U.S. is \$4, arbitrageurs will try to gain by exploiting these price differences. If they purchase the commodity in the U.S., transport it at zero cost to the U.K., and sell it in the U.K., then the rising demand in the U.S. and the rising supply in the U.K. will ensure that the prices become equal. For example, let's suppose that the price in the U.S. rises to \$5 per unit and the price in the U.K. falls to £2.50 per unit. Once we convert pounds into dollars in the U.K., we can see that the U.K. price is now \$5, which is the same as the U.S. price.

The analysis we just completed sidestepped a very important issue. That is, when international arbitrageurs demand more of the low-priced U.S. commodity, they must first buy dollars in the foreign exchange market. They then purchase the commodity in the U.S. using dollars and transport the commodity to the U.K. where they sell it for pounds at a high price. The pounds are then sold in the foreign exchange market for dollars, which are then used to buy the commodity in the U.S. again, and the cycle repeats until the prices equalize. It is important to notice that the purchase of dollars and the sale of pounds in the foreign exchange market will cause the dollar-pound exchange rate to change under a floating exchange rate regime. That is, the dollar will appreciate, and the pound will depreciate.

The adjustments that take place in the foreign exchange market are very rapid and so let's assume that these adjustments occur before the U.S. demand for the product and the U.K. supply of the product have a chance to change. Suppose that the pound-dollar exchange rate becomes £0.75/$ and the dollar-pound exchange rate becomes $1.3333/£ (i.e., 4/3 pounds to be exact). Now the U.K. price is equal to $4 (= £3 times $1.3333/£). That is, the prices have equalized, and they became equal entirely due to exchange rate adjustments, as shown in Figure 20.18.

Figure 20.18: A Case of International Arbitrage (assuming flexible exchange rates that adjust from $2/£ to $1.3333/£)



For commodities that are easily tradeable across nations, we should expect an outcome like what we have just described. For commodities that are not easily traded (e.g., services like haircuts), we should expect to observe more variation in purchasing power. That is, we should expect the prices to be quite different after we make the conversion into a common currency. Therefore, while this theory provides a useful starting point for thinking about international price adjustments, limits to its application exist, particularly where excessive transport costs interfere with arbitrage.

Marxist Approaches to Imperialist Finance

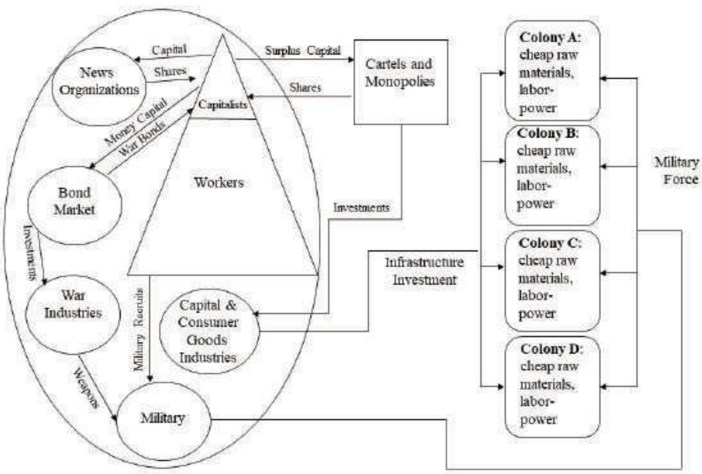
In this section, we will consider Marxist approaches to foreign exchange markets and imperialist finance. Theories of **imperialism** have a long history in Marxian economics, stretching back to the works of Rudolf Hilferding, J.A. Hobson, Rosa Luxemburg, and V.I. Lenin in the early part of the twentieth century.[10] Because these different theories of imperialism have much in common, this section provides a brief overview of Hobson's theory of imperialism, which emphasizes several key aspects of this body of thought.

This section draws upon E.K. Hunt's account of Hobson's theory from his valuable *History of Economic Thought* (2002).[11] During the late nineteenth century, many of the least developed places on Earth were colonized as the capitalist powers grabbed colonial possessions in Africa and Asia.[12] According to Hobson, the many reasons given for overseas military adventures by the advanced capitalist nations during this period had little to do with the publicly stated reasons that were offered at the time. The suggestion that the primary purpose was to spread Christianity to uncivilized people in backward lands was a distortion that hid the true motives of the imperialist nations.[13] Hobson dismissed other suggestions about the root cause of imperialism, including the claim that militaristic tendencies were an inherent part of human nature.[14] After all, the surge of imperialism in the late nineteenth century must have a cause that is connected to historical circumstances. Hobson also rejected the notion that the imperialist aggression of the period simply stemmed from the "irrational nature" of politicians.[15] In Hobson's view, even though the activities might appear irrational from the perspective of the nation, they benefitted certain classes within the nation a great deal.[16]

Hobson's explanation for the imperialist tendencies of the late nineteenth century concentrates on the massive accumulation of capital that became concentrated in the hands of large banks and financial houses during the late nineteenth century.[17] So much capital had accumulated and the disparity between workers and capitalists had become so great that profitable domestic outlets for the surplus capital could not be found even as capitalists spent enormous amounts on luxuries.[18] As a result, financial capitalists began to look elsewhere for a way to invest surplus capital and prop up the demand for commodities.

The capitalist classes in the imperialist nations found a new way to relieve their economic problems at home. Financial capitalists bought government bonds, which helped finance military production. They also bought shares to help finance the activities of international cartels and global monopolies. These activities allowed large corporations to invest in production in the nation's colonial possessions. To carry on this overseas production, it was necessary to develop the infrastructure in the colonies, which it achieved by purchasing capital goods from the imperialist nation.[19] The expanded infrastructure in the colonies created a large network of roads, bridges, and railroads that made it possible to transform pre-capitalist societies into capitalist societies.[20] These changes made it possible to acquire vast possessions of cheap raw materials and transport them for use in capitalist production using the cheap wage labor of the colonies.[21] The increase in the wage labor forces in the colonies also created a large demand for the consumer goods of the imperialist nations,[22] which helped alleviate the insufficient aggregate demand problem at home and which boosted the profits of the global capitalist enterprises. These relationships are captured in Figure 20.19.

Figure 20.19: An Imperialist Road Map



To make all this imperialist activity possible, it was necessary to promote feelings of patriotism and nationalism within the population of the imperialist nation.[23] That is, the public needed to be convinced of the righteousness of the imperialist cause. That required capitalist support for pro-war messages in the newspapers. With enough willing military recruits, the wars to seize and hold colonial possessions could be fought and won, as shown in Figure 20.19. Without this sort of imperialist activity, the result would be more severe business cycles and depressions in the home nation because capitalists would fail to find profitable outlets for the surplus capital they controlled.[24]

Although Hobson's theory applied to late nineteenth century imperialism, it can be modified easily to apply to the present neo-colonial period in which the capitalist nations use military aggression in support of regimes in developing nations. Such regimes are willing to grant the imperialist nations production contracts and access to cheap raw materials and labor-power. In return for their cooperation, the imperialist power provides infrastructure development, including oil pipelines and military bases to protect the regimes of the subjugated nations. These regimes are frequently puppet governments that operate under the guidance of the imperialist nation.

A Marxian Approach to Exchange Rate Determination

Because much of this chapter has focused on the determination of exchange rates and the factors that influence them, it is worth considering what Marxian economists might have to say about exchange rates. In Chapter 17, a Marxian theory of fiat money was developed. In that chapter, it was shown that Marxian economists can explain the determination of the value of fiat money by dividing the product of the money supply and the velocity of money in circulation (MV) by the aggregate labor time embodied in the circulating commodities (L) during a given period. The calculation produces the **monetary expression of labor time** (MELT), which is in units of currency per hour of socially necessary abstract labor time (e.g., $16 per hour).

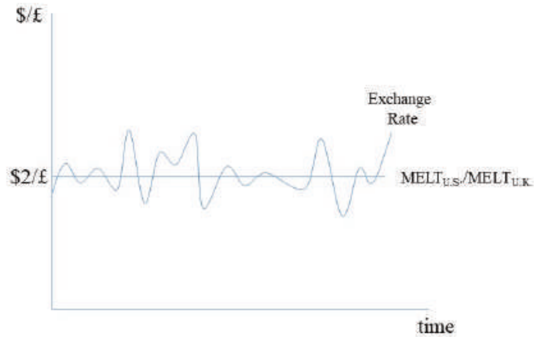
The MELT makes it possible to convert any labor time magnitude into its monetary equivalent.

Any nation taken by itself will have a MELT in terms of its home currency per unit of socially necessary abstract labor time. It is relatively easy to determine the foreign exchange rate between two currencies if you happen to know the MELTs in each nation. For example, suppose that the MELT in the United States is $16 per hour and the MELT in the United Kingdom is £8 per hour. The dollar-pound exchange rate (e), or the **foreign exchange value** of the pound can then be calculated as follows:

$$e = \frac{MELT_{U.S.}}{MELT_{U.K.}} = \frac{\$16/hour}{£8/hour} = \frac{\$2}{£}$$

In theory, we should expect this exchange rate to emerge in the international marketplace. However, deviations from this value are to be expected. In Chapter 4, it was pointed out that supply and demand can cause deviations of price from commodity value in Marxian economics. The same applies in the foreign exchange market. A change in the supply or demand of pounds (dollars) can cause the actual **market exchange rate** to deviate from its expected value. Such fluctuations may be rather large, but over time exchange rates will gravitate towards their MELT-determined values, as shown in Figure 20.20.

Figure 20.20: The Influence of Supply and Demand on Exchange Rates in Marxian Economics



It should be noted that Marxian economists are interested in understanding the social relationships that give rise to the appearances they observe in the marketplace. This theory of exchange rate determination should be viewed only as a complement to the theory of imperialist finance that was developed in the previous section. The class conflict inherent in the vast network connecting the imperialist power with its colonies (or neo-colonies) is what ultimately concerns the Marxist. It is the focus because a solid understanding of it will assist in the revolutionary overthrow of the capitalist economic system that makes such relations possible.

Following the Economic News[25]

According to an article that appeared in *The Wall Street Journal,* the Mexican peso hit an all-time low against the dollar in July of 2015 at about 16 pesos per dollar. Other Latin American currencies experienced significant declines. The Colombian peso and the Brazilian real were at their lowest levels relative to the U.S. dollar in more than 10 years. The Peruvian sol and the Chilean peso both also suffered reductions in value relative to the dollar. What was happening that caused such steep currency depreciations in these countries? One major reason is that many investors expect the U.S. Federal Reserve to raise interest rates relatively soon. As we have learned, an increase in U.S. interest rates relative to interest rates in other nations will tend to cause an appreciation of the dollar and a depreciation of other currencies, other factors the same. The Bank of Mexico, the Mexican central bank, is expected to raise interest rates when the Fed does, and has already been selling off dollars to maintain the value of the peso. Nevertheless, the peso has continued to depreciate, in part due to falling oil prices, which dampen the demand for the peso. Still, the weaker peso might give a boost to Mexico's export markets. As the author explains, the weaker peso also raises the local value of funds sent by Mexicans working in the U.S. to their families in Mexico. Slower global economic growth has had an impact on commodity prices, generally, which has helped to depress several currencies. The Colombian peso has depreciated in part due to lower oil prices. The Chilean peso has depreciated with the drop in the price of copper as Chinese copper demand has declined. The Peruvian sol has also dropped in value as the demand for its currency has declined along with the reduction in demand for its exports of copper, gold, zinc, and lead. The supply and demand model of the foreign exchange market that we developed in this chapter provides a helpful framework that can be easily applied to understand changes in exchange rates. In this case, changes in relative interest rates and changes in income levels due to slowing global economic growth appear to be the primary causes of the exchange rate movements.

Summary of Key Points

1. In the balance of payments accounts, each payment that a nation receives from the rest of the world is recorded as a credit and each payment that a nation makes to the rest of the world is recorded as a debit.
2. The current account is a subaccount of the balance of payments statement that records transactions that do not involve the purchase and sale of income-earning assets.
3. The capital account is a subaccount of the balance of payments statement that records transactions involving non-financial assets, whereas the financial account records transactions involving financial assets.
4. If a current account surplus exists, then a capital and financial account deficit of the same amount must exist. If a current account deficit exists, then a capital and financial account surplus of the same amount must exist.
5. When a nation's net international investment position is positive (its foreign assets exceeds its foreign liabilities), then it is a net creditor. When a nation's net international investment position is negative (its foreign assets fall short of its foreign liabilities), then it is a net debtor.
6. A nation's current account balance may be expressed as the sum of its private sector gap and its government budget gap.
7. If a nation's currency appreciates relative to

another currency, then the other currency depreciates relative to the nation's currency.

8. Changes in the quantity demanded or quantity supplied of a foreign currency may only be caused by a change in the exchange rate. If any other factors change that might affect buyers and sellers of a foreign currency, then changes in demand and supply occur.
9. When potential balance of payments deficits exist, official reserves are used to guarantee an overall payments balance. When potential balance of payments surpluses exist, official reserves are accumulated to guarantee an overall payments balance.
10. International arbitrage ensures that exchange rates adjust to equalize the purchasing power of different currencies.
11. Marxist theories of imperialism emphasize capitalists' inability to find profitable domestic outlets for surplus capital as the reason for foreign conquest and the subjugation of foreign people in less developed parts of the world.
12. Marxian economists explain the average level of a foreign exchange rate over time using the ratios of the monetary expressions of labor time (MELTs) of two different nations.

List of Key Terms

Balance of payments accounting

National income accounting

Credit

Debit

Current account

Current account balance

Current account surplus

Current account deficit

Trade balance

Trade deficit

Trade surplus

Balanced trade

Capital account

Financial account

Official reserve holdings

Capital and financial account surplus

Capital and financial account deficit

Statistical discrepancy

Net investment position

Net International Investment Position

Net debtor

Net creditor

Depreciated

Appreciated

Equilibrium exchange rate

Equilibrium quantities exchanged

Capital gain

Capital loss

Change in quantity demanded

Change in demand

Change in quantity supplied

Change in supply

Self-fulfilling prophecy

Exchange rate regime

Floating exchange rate regime

Fixed exchange rate regime

Devaluation

Revaluation

Managed floating exchange rates

Currency crisis

Theory of purchasing power parity

Arbitrage

Imperialism

Monetary expression of labor time (MELT)

Foreign exchange value

Market exchange rate

Problems for Review

1. Suppose that a current account deficit of $400 billion exists, that a capital account surplus of $100 billion exists, and that statistical discrepancy is $75 billion. What is the state of the financial account?
2. Suppose that the net international investment position of the U.S. at the start of 2014 is negative $7.1 trillion. Also assume that a current account surplus of $400 billion exists at the end of 2014. What is the net international investment position at the end of 2014?
3. Suppose that the Yen-U.S. dollar exchange rate is ¥110/$. What is the U.S. dollar-Yen exchange rate in this case?
4. Suppose that the Yen-U.S. dollar exchange rate changes from ¥110/$ to ¥115/$. What has happened to the foreign exchange value of the U.S. dollar? What has happened to the foreign exchange value of the Yen?
5. Suppose that the incomes rise significantly in India with conditions in the rest of the world remaining roughly the same. What will happen in the U.S. dollar-rupee market? Analyze the situation using the three steps shown in Figures 20.9 and 20.10. Only consider the sellers' side of the market for rupees in your answer.
6. Suppose that interest rates fall in Europe with

conditions in the rest of the world remaining roughly the same. What will happen in the U.S. dollar-Euro market? Analyze the situation using the three steps shown in Figures 20.9 and 20.10. Only consider the buyers' side of the market for Euros in your answer.

7. Suppose that an automobile costs $12,000 in the United States and ¥1,300,000 in Japan. If the Yen-U.S. dollar exchange rate is ¥100/$, then does purchasing power parity hold in this case? If not, then what needs to happen to the dollar (appreciate or depreciate) for purchasing power parity to hold?
8. Suppose that the MELT in the United States is $22 per hour of socially necessary abstract labor time (SNALT) and the MELT in Russia is 0.3793 rubles per hour of SNALT. What is the foreign exchange value of the ruble in this case?

Notes

1. See Bureau of Economic Analysis, U.S. International Transactions Accounts Data, at www.bea.gov.
2. See Bureau of Economic Analysis, U.S. International Transactions Accounts Data, at www.bea.gov.
3. The loose inspiration for this brief section stems from Melvin (2000), p. 1-6.
4. The explanations of the shapes of the demand and supply curves of foreign exchange are found in many introductory neoclassical textbooks. For example, see Bade and Parkin (2013), p. 880-884, and McConnell and Brue (2008), p. 700-701.

5. Bade and Parkin (2013), p. 880-884, refer to the expected profit effect when discussing this effect in their discussion of the derivation of the demand and supply curves in the foreign exchange market.
6. Lists of the factors discussed in this section (that affect both the supply and demand sides of the market) are commonly found in neoclassical textbooks. For example, McConnell and Brue (2008), p. 701-703, analyze the same factors. Carbaugh (2011), p. 407-421, analyzes the same factors except that he emphasizes productivity differences rather than income differences. Carbaugh (2011), p. 410, also mentions trade barriers as a factor.
7. As Samuelson and Nordhaus (2001), p. 622, explain, movements of exchange rates serve as "a balance wheel to remove disequilibria in the balance of payments."
8. Stiglitz (2002), p. 89.
9. Ibid. p. 94-95.
10. See Hunt (2002), p. 351.
11. Ibid. p. 351-356.
12. Hunt (2002), p. 348-351, provides an overview of this period.
13. Ibid. p. 351.
14. Ibid. p. 352.
15. Ibid. p. 352.
16. Ibid. p. 352.
17. Ibid. p. 353.
18. Ibid. p. 354.
19. Some of these elements are given greater emphasis in Rosa Luxemburg's account of imperialism. See Hunt (2002), p. 361.
20. Ibid. p. 361.
21. Ibid. p. 359.

22. Ibid. p. 361.
23. Hobson argues that finance "manipulates the patriotic forces" of the population. See Hunt (2002), p. 354.
24. Ibid. p. 356.
25. Harrup, Anthony. "Latin American Currencies Hurt by Commodities' Drop, U.S. Fed Expectations." *The Wall Street Journal*. July 20, 2015.

SELECTED REFERENCES

Antunes de Oliveira, Felipe. "A Radical Invitation for Latin America." *Monthly Review*. May 1, 2017. Web. Accessed on March 26, 2019. https://monthlyreview.org/2017/05/01/a-radical-invitation-for-latin-america/

Armentano, D.T. "A Critique of Neoclassical and Austrian Monopoly Theory." In *New Directions in Austrian Economics*. Edited by Louis M. Spadaro. Institute for Humane Studies: Menlo Park, CA; 1978.

Aslaksen, Iulie. "Gross Domestic Product." *The Elgar Companion to Feminist Economics*. Ed. Janice Peterson and Margaret Lewis. Edward Elgar. Northampton, MA. 1999.

Bade, Robin and Michael Parkin. *Foundations of Economics*. Sixth Edition. Pearson: New York, 2013.

Bagdikian, Ben H. *The New Media Monopoly*. Beacon Press: Boston, MA; 2004.

Brewer, Anthony. *Marxist Theories of Imperialism: A Critical Survey*. Second Edition. Routledge: New York, 1990.

Carbaugh, Robert J. *International Economics*. 13th Edition. South-western Cengage Learning: Mason, OH: 2011.

Carlton, Dennis W. and Jeffrey M. Perloff. *Modern*

Industrial Organization. Third Edition. Addison-Wesley: New York, 2000.

Carrasco, Enrique R. "The 1980s: The Debt Crisis & The Lost Decade." *The E-Book on International Finance & Development*. The University of Iowa Center for International Finance and Development. June 6, 2011. Web. Accessed on April 27, 2018. https://web.archive.org/web/20110606041529/http://www.uiowa.edu/ifdebook/ebook2/contents/part1-V.shtml

Case, Karl E., Ray C. Fair, and Sharon M. Oster. *Principles of Economics*. Tenth Edition. Prentice Hall: New York, 2012.

Chancellor, Edward. *Devil Take the Hindmost: A History of Financial Speculation*. Plume: New York, 1999.

Chiang, Eric P. and Gerald W. Stone. *Core Economics*. Third Edition. Worth Publishers: New York, 2014.

Chisholm, Roger and Marilu McCarty. *Principles of Macroeconomics*. Second Edition. Scott, Foresman and Company: Glenview, IL, 1981.

Chomsky, Noam. *Understanding Power*. The New Press: New York, 2002.

Coppock, Lee and Dirk Mateer. *Principles of Macroeconomics*. W.W. Norton: New York, 2014.

Coppock, Lee and Dirk Mateer. *Principles of Microeconomics*. W.W. Norton: New York, 2014b.

Desai, Meghnad. *Marx's Revenge*. Verso: New York, 2004.

Eatwell, John. *The New Palgrave: Marxian Economics*. Edited by John Eatwell, Murray Milgate, and Peter Newman. W.W. Norton: New York, 1990.

Emmanuel, Arghiri. *Unequal Exchange: A Study of the Imperialism of Trade*. Monthly Review Press: New York, 1972.

Fletcher, Ian. *Free Trade Doesn't Work: Why America Needs a Tariff*. U.S. Business & Industry Council: Washington, DC, 2009.

Foster, John Bellamy and Fred Magdoff. *The Great Financial Crisis: Causes and Consequences*. Monthly Review Press: New York, 2009.

Frank, Andre Gunder. *Capitalism and Underdevelopment in Latin America*. Monthly Review Press: New York, 1969.

Garrison, Roger W. "Austrian Macroeconomics: A Diagrammatical Exposition." *New Directions in Austrian Economics*. Ed. Louis M. Spadaro. Sheed Andrews and McMeel: Kansas City, MO; 1978.

Gleicher, David. "A Novel Method of Teaching Keynesian Demand versus Neoliberal Austerity Policies within the Simple Keynesian Model." Union for Radical Political Economics (URPE) Newsletter. Volume 44, Number 3. Spring 2013.

Gomory, Ralph E. and William J. Baumol. *Global Trade and Conflicting National Interests*. The MIT Press: Cambridge, MA; 2000.

Greenspan, Alan. "Antitrust." In *Capitalism: The Unknown*

Ideal by Ayn Rand. New American Library: New York, 1967.

Hessen, Robert. *Steel Titan: The Life of Charles M. Schwab*. Oxford University Press: New York, 1975.

Howard, M.C. and J.E. King. *A History of Marxian Economics: Volume II, 1929-1990.* Princeton University Press: Princeton, NJ; 1992.

Hubbard, R. Glenn and Anthony Patrick O'Brien. *Economics*. Seventh Edition. Pearson: New York, 2019.

Hunt, E.K. and Mark Lautzenheiser. *History of Economic Thought: A Critical Perspective*. Third Edition. M.E. Sharpe: New York, 2011.

Hunt, E.K. *History of Economic Thought: A Critical Perspective*. Updated Second Edition. M.E. Sharpe: Armonk, New York. 2002.

Itoh, Makoto. "Unequal Exchange Reconsidered in our Age of Globalization." *Heterodox Macroeconomics: Keynes, Marx and globalization*. Edited by Jonathan P. Goldstein and Michael G. Hillard. Routledge: New York, 2009.

Keat, Paul G., Philip K.Y. Young, and Stephen E. Erfle. *Managerial Economics*. Seventh Edition. Pearson: New York, 2013.

Kenyon, Peter. "Pricing." *A Guide to Post-Keynesian Economics*. Edited by Alfred S. Eichner. M.E. Sharpe: White Plains, NY; 1978.

Krugman, Paul, Robin Wells, and Kathryn Graddy.

Essentials of Economics. Third Edition. Worth Publishers: New York, 2014.

Lianos, Theodore. "Marx and the Rate of Interest." *Review of Radical Political Economics*. Volume 19 (3): 34-55, 1987.

Lipsey, Richard G. and Paul N. Courant. *Microeconomics*. Eleventh Edition. HarperCollins College Publishers: New York, 1996.

Mankiw, N. Gregory. *Macroeconomics*. Worth Publishers: New York, 1997.

Marx, Karl. *Capital: A Critique of Political Economy. Volume 2.* Penguin Group: New York, 1992.

Marx, Karl. *Capital: A Critique of Political Economy*. *Volume 3.* Penguin Group: New York, 1991.

Marx, Karl. *Capital: A Critique of Political Economy*. *Volume 1.* Penguin Group: New York, 1990.

McConnell, Campbell R., Stanley L. Brue, and Sean M. Flynn. *Economics*. Instructor's Edition. McGraw-Hill Irwin: New York, 2012.

McConnell, Campbell R. and Stanley L. Brue. *Economics: Principles, Problems, and Policies*. Seventeenth Edition. McGraw-Hill Irwin: New York, 2008.

McConnell, Campbell R. and Stanley L. Brue. *Macroeconomics*. Thirteenth Edition. McGraw-Hill: New York, 1996.

McDonough, Terrence. "The Construction of Social Structures of Accumulation in U.S. History" in

Social Structures of Accumulation. Ed. David Kotz et al. Cambridge University Press: New York, NY. 1994.

Melvin, Michael. *International Money and Finance*. Sixth Edition. Addison-Wesley: New York, 2000.

Mirowski, Philip. *More Heat Than Light*. Cambridge University Press: New York, 1991.

Mishkin, Frederic S. *The Economics of Money, Banking, and Financial Markets*. Seventh Edition Update. Pearson: New York, 2006.

Moseley, Fred. "The Determination of the "Monetary Expression of Labor Time" ("MELT") in the Case of Non-Commodity Money. *Review of Radical Political Economics*. 43 (1). 2011. p. 95-105.

Moseley, Fred. "Introduction." *Marx's Theory of Money: Modern Appraisals*. Edited by Fred Moseley. Palgrave Macmillan: New York, 2005. p. 1-18.

Moseley, Fred. "The Monetary Expression of Labor": in the Case of Non-commodity Money." November 2004. Web. https://www.mtholyoke.edu/~fmoseley/

Moss, Laurence. "The Emergence of Interest in a Pure Exchange Economy: Notes on a Theorem Attributed to Ludwig von Mises." In *New Directions in Austrian Economics,* edited by Louis Spadaro. Sheed Andrews and McMeel: Kansas City, MO; 1978.

Oakland, William H. "Budgetary Measures of Fiscal Performance." *Southern Economic Journal,* Vol. 35, No. 4, April 1969, pp. 347-358.

OpenStax. *Principles of Economics*. Rice University: Houston, TX; 2014. https://openstax.org/details/books/principles-economics

Pindyck, Robert S. and Daniel L. Rubinfeld. *Microeconomics*. Eighth Edition. Pearson: New York, 2013.

Rand, Ayn. "Appendix: Man's Rights." *Capitalism: The Unknown Ideal*. New American Library: New York, 1966.

Rose, Matthew. "Andre Gunder Frank & Dependency Theory." *Our Politics*. September 9, 2016. Web. Accessed on March 26, 2019. https://ourpolitics.net/andre-gunder-frank-dependency-theory/

Rosser, J. Barkley, Jr. and Marina V. Rosser. *Comparative Economics in a Transforming World Economy*. Second Edition. The MIT Press: Cambridge, MA; 2004.

Rothbard, Murray. "Austrian Definitions of the Supply of Money." In *New Directions in Austrian Economics*. Ed. Louis M. Spadaro. Sheed Andrews and McMeel: Kansas City, MO; 1978.

Solomon, Robert. "A Note on the Full Employment Budget Surplus." *The Review of Economics and Statistics*, Volume 46, No. 1, February 1964, pp. 105-108.

Samuelson, Paul A. and William D. Nordhaus. *Economics*. Seventeenth Edition. McGraw-Hill Irwin: New York, 2001.

Saros, Daniel E. "The Circulation of Bank Capital and the General Rate of Interest." *Review of Radical Political Economics*. 45 (2). Spring 2013.

Saros, Daniel E. "The Price-Form as a Fractional Reflection of the Aggregate Value of Commodities. *Review of Radical Political Economics*. 39 (3). Summer 2007. p. 407-415.

Saros, Daniel E. "The Price-Form as a Fractional Reflection of the Aggregate Value of Commodities (2006)." July 4, 2006. Web. Available at: https://works.bepress.com/daniel-saros/11/

Saros, Daniel E. "The Price-Form as a Fractional Reflection of the Aggregate Value of Commodities (2002)." May 5, 2002. Web. Available at: http://works.bepress.com/daniel-saros/8/.

Shaikh, Anwar. "The Laws of International Exchange." *Growth, Profits and Property: Essays in the Revival of Political Economy*. Edited by Edward J. Nell. Cambridge University Press: Cambridge, MA; 1980. p. 204-235.

Snowdon, Brian, Howard Vane, and Peter Wynarczyk. *A Modern Guide to Macroeconomics: An Introduction to Competing Schools of Thought*. Edward Elgar: Northampton, MA; 1994.

Stiglitz, Joseph E. *Globalization and Its Discontents*. W.W. Norton: New York, 2002.

U.S. Bureau of the Census. *Historical Statistics of the United States, Colonial Times to 1970, Bicentennial Edition, Part 2*. Washington, DC; 1975.

Varian, Hal R. *Intermediate Microeconomics: A Modern Approach*. Fifth Edition. W.W. Norton & Company. New York, 1999.

Wolff, Leon. *Lockout: The Story of the Homestead Strike of 1892 – A Study of Violence, Unionism, and the Carnegie Steel Empire*. Harper & Row: New York, 1965.

Wolff, Richard. "Capitalism's Crisis through a Marxian Lens." *Capitalism Hits the Fan*. Olive Branch Press: Northampton, MA; 2010.

Wolff, Richard and Stephen Resnick. *Contending Economic Theories: Neoclassical, Keynesian, and Marxian*. The MIT Press: Cambridge, MA; 2012.

Wolfson, Martin H. "The Financial System and the Social Structure of Accumulation" in *Social Structures of Accumulation*. Ed. David Kotz et al. Cambridge University Press: New York, NY. 1994.