CONTENTS

Course Contents
- About This Course
- Learning Outcomes

Module 1: Economic Thinking
- Why It Matters: Economic Thinking
- Introduction to Economics
- Understanding Economics and Scarcity
- The Concept of Opportunity Cost
- Labor, Markets, and Trade
- Microeconomics and Macroeconomics
- Introduction to Math in Economics
- Economic Models
- Purpose of Functions
- Solving Simple Equations
- Introduction to Graphs in Economics
- Creating and Interpreting Graphs
- Interpreting Slope
- Types of Graphs
- Putting It Together: Economic Thinking

Module 2: Choice in a World of Scarcity
- Why It Matters: Choice in a World of Scarcity
- Introduction to the Cost of Choices
- Budget Constraints and Choices
- Calculating Opportunity Cost
- Learn By Doing: Budget Constraints and Opportunity Cost
- Introduction to the Production Possibilities Frontier
- The Production Possibilities Frontier
- Productive Efficiency and Allocative Efficiency
- Introduction to the Economic Way of Thinking
- Rationality and Self-Interest
- Marginal Analysis
- Positive and Normative Statements
- Putting It Together: Choice in a World of Scarcity

Module 3: Supply and Demand
- Why It Matters: Supply and Demand
- Introduction to Economic Systems
- Economic Systems
- Introduction to Demand
- What Is Demand?
- Factors Affecting Demand
- Learn By Doing: Graphing Demand
- Learn By Doing: Demand for Food Trucks
- Introduction to Supply
- What is Supply?
- Factors Affecting Supply
- Learn By Doing: Graphing Supply and Demand
- Learn By Doing: Supply of Food Trucks
- Introduction to Equilibrium
- Equilibrium, Surplus, and Shortage
- Changes in Equilibrium
- Finding Equilibrium
- Changes in Supply and Demand
- Learn By Doing: Food Trucks and Changes in Equilibrium
- Learn By Doing: Calculating Equilibrium
Putting It Together: Supply and Demand

Module 4: Applications of Supply and Demand
- Why It Matters: Applications of Supply and Demand
- Introduction to Price Ceilings and Price Floors
- Price Ceilings
- Price Floors
- Learn By Doing: Shortage and Surplus
- A Closer Look at Price Controls
- Introduction to Surplus
- Trade and Efficiency
- Consumer & Producer Surplus
- Inefficiency of Price Floors and Price Ceilings
- Learn By Doing: Consumer and Producer Surplus
- Introduction to Labor and Financial Markets
- Labor and Financial Markets
- Putting It Together: Applications of Supply and Demand

Module 5: Elasticity
- Why It Matters: Elasticity
- Introduction to Elasticity
- Elasticity of Demand
- Examples of Elastic and Inelastic Demand
- Introduction to Calculating Price Elasticity
- Calculating Elasticity and Percentage Changes
- Calculating Price Elasticities Using the Midpoint Formula
- Learn By Doing: Calculating Price Elasticities
- Categories of Elasticity
- Price Elasticity of Supply
- Introduction to Elasticities in Areas Other Than Price
- Income Elasticity, Cross-Price Elasticity & Other Types of Elasticities
- Introduction to Price Elasticity and Total Revenue
- Elasticity and Total Revenue
- Elasticity, Costs, and Customers
- Tax Incidence
- Putting It Together: Elasticity

Module 6: Utility
- Why It Matters: Utility
- Introduction to Utility and Consumer Equilibrium
- Consumer Choice and Utility
- Marginal Utility versus Total Utility
- Rules for Maximizing Utility
- Learn By Doing: Maximizing Utility
- Introduction to Changes in Consumer Equilibrium
- Income Changes and Consumption Choices
- The Foundations of the Demand Curve
- Introduction to Indifference Curves
- Indifference Curve Analysis
- Introduction to Behavioral Economics
- Behavioral Economics: An Alternative Viewpoint
- Putting It Together: Utility

Module 7: Production and Costs
- Why It Matters: Production and Costs
- Introduction to Production
- What is Production?
- The Production Function
- Introduction to Costs in the Short Run
- Costs and Profit
- Costs in the Short Run
Module 8: Perfect Competition
- Why It Matters: Perfect Competition
- Introduction to Perfect Competition
- Perfect Competition
- Profit Maximization in a Perfectly Competitive Market
- Learn By Doing: Profit Maximization in a Perfectly Competitive Firm
- Introduction to Profit in a Perfectly Competitive Firm
- Calculating Profits and Losses
- The Shutdown Point
- Introduction to the Long Run and Efficiency in Perfectly Competitive Markets
- Entry and Exit Decisions in the Long Run
- Learn By Doing: Entry and Exit Decisions in the Long Run
- Efficiency in Perfectly Competitive Markets
- Putting It Together: Perfect Competition

Module 9: Monopoly
- Why It Matters: Monopoly
- Introduction to Monopoly
- Monopolies
- How Monopolies Form: Barriers to Entry
- Introduction to Profit and Losses in Monopolies
- Profit Maximization for a Monopoly
- Learn By Doing: Profit Maximization for a Monopoly
- Computing Monopoly Profits
- The Inefficiency of Monopoly
- Price Discrimination and Efficiency
- Introduction to Monopoly and Antitrust Policy
- Corporate Mergers
- Regulating Anticompetitive Behavior
- Regulating Natural Monopolies
- The Great Deregulation Experiment
- Putting It Together: Monopolies

Module 10: Monopolistic Competition and Oligopoly
- Why It Matters: Monopolistic Competition and Oligopoly
- Introduction to Monopolistically Competitive Industries
- Monopolistic Competition
- Introduction to Analyzing and Graphing Monopolistic Competition
- Profit Maximization under Monopolistic Competition
- Learn By Doing: Profit Maximization Under Monopolistic Competition
- Entry, Exit and Profits in the Long Run
- Monopolistic Competition and Efficiency
- Product Differentiation and Advertising
- Introduction to Oligopolies
- Why do Oligopolies Exist?
- Collusion or Competition?
- Prisoner's Dilemma
- Learn By Doing: Prisoner's Dilemma
- Putting It Together: Monopolistic Competition and Oligopoly
Module 15: Exchange Rates and International Finance
- Why It Matters: Exchange Rates and International Finance
- Introduction to the Foreign Exchange Market
- The Foreign Exchange Market
- Strengthening and Weakening Currency
- Introduction to Exchange Rates and Purchasing Power
- Demand and Supply Shifts in Foreign Exchange Markets
- Introduction to Exchange Rates and the Trade Balance
- Macroeconomic Effects of Exchange Rates
- Exchange-Rate Policies
- Putting It Together: Exchange Rates and International Finance
Waymaker Microeconomics (Summer 2018)

This course provides a strong foundation in microeconomic theory, whether preparing for further study in economics, the social sciences, business, or other disciplines. Designed to help students think like economists, course materials use engaging, real-world examples to explore how individuals and firms make economic choices. Key topics include supply and demand, elasticity, utility, production and costs, and an analysis of types of markets: perfect competition, monopoly, monopolistic competition, and oligopoly. Content coverage also includes public goods and externalities, labor markets, income distribution, globalization and trade, and exchange rates and international finance. An orientation module helps students refresh skills around modeling, graphing, and algebra for economic problem-solving.

Contributors

This course, based on OpenStax Principles of Economics, includes additional noteworthy contributions by the Lumen Learning team and:

- Steve Greenlaw (University of Mary Washington)
- Sophie Haci (Houston Community College)
- Melissa Walker (Nashville State Community College)
- Veronika Dolar (SUNY Old Westbury)
- Michael Fusillo (Tufts University)
- Shawn Kilpatrick (Northeastern University)
- Jennifer Pakula (Saddleback College)
- Clark Aldrich

What’s New?

This new edition of Waymaker Microeconomics includes significant improvements and enhancements:

Data-Driven Improvements

Several topics proved to be consistently difficult for the thousands of students who took a Waymaker Economics course during Fall 2017. We have made specific improvements to address these difficult topics, several of which are explained below:

- The Production Possibilities Frontier: The new version of the course includes additional applied graphing practice and two new videos on this topic (see https://courses.lumenlearning.com/wm-microeconomics/chapter/the-production-possibilities-frontier/)
- Price Elasticity and Total Revenue: The new version of the course includes additional applied practice on this topic (see https://courses.lumenlearning.com/wm-microeconomics/chapter/elasticity-and-total-revenue/)
- Surplus and Efficiency: The new version of the course includes a “Learn by Doing” page for additional practice in solving for consumer and producer surplus (see https://courses.lumenlearning.com/wm-
microeconomics/chapter/learn-by-doing-consumer-and-producer-surplus/

- Graphs in Economics: The new version of the course includes several new practice questions, including those which enable students to manipulate graphs (see https://courses.lumenlearning.com/wm-microeconomics/chapter/interpreting-slope/)

General Improvements

- **Improved course organization**
  - Modules are more consistent in size and depth
  - Content is organized around specific, granular, learning outcomes, which are listed at the top of each page
  - Practice questions, try it questions, and quiz questions all align with learning outcomes
- **“Try It” embedded practice questions** for every learning outcome (for example, on this page, Factors Affecting Demand, students learn about concepts and then immediately check their understanding with applied practice)
  - Some quantitative “Try It” questions are intentionally designed for unlimited practice using different number sets (example at the bottom of this page on Price Floors)
- **“Learn By Doing” pages** for even more practice
  - Similar to the “Try It” questions, some “Learn By Doing” pages consist of quantitative practice questions that serve as another checkpoint for students to assess their own understanding (see Learn By Doing: Graphing Demand)
  - Other “Learn By Doing” pages are simulations which allow students to manipulate variables and assess the impact of decisions (see Learn By Doing: Demand for Food Trucks)
- **“Watch It” embedded videos** that explain and reiterate key concepts throughout the course
  - Most videos come from Marginal Revolution University, Dr. Mary McGlasson (mjmfoodie), ACDC Economics, and Crash Course Economics
- **Interactive graphs** (as seen in the salmon fishing example on this page: Changes in Equilibrium)
  - Complicated graphs are explained in a step-by-step process to help students see and understand how the graph changes
- **Discussions and Assignments** for every module (available here with instructor log-in)
- **Problem Sets** for every module (available here with instructor log-in)

About Lumen

Lumen Learning’s mission is to enable unprecedented learning for all students.

We do this by using open educational resources (OER) to create well-designed and low-cost course materials that replace expensive textbooks. Because learning is about more than affordability and access, we also apply learning science insights and efficacy research to develop learning activities that are engineered to improve subject mastery, course completion and retention.

If you’d like to connect with us to learn more about adopting this course, please Contact Us.

You can also make an appointment for OER Office Hours to connect virtually with a live Lumen expert about any question you may have.
The content, assignments, and assessments for this course are aligned to the following learning outcomes. A full list of course learning outcomes can be viewed here: Microeconomics Course Learning Outcomes.

Module 1: Economic Thinking
- Explain what economics is and explain why it is important
- Use mathematics in common economic applications
- Use graphs in common economic applications

Module 2: Choice in a World of Scarcity
- Explain the cost of choices and trade-offs
- Illustrate society’s trade-offs by using a production possibilities frontier, or curve
- Explain the assumption of rationality by individuals and firms

Module 3: Supply and Demand
- Describe and differentiate between major economic systems
- Explain the determinants of demand
- Explain the determinants of supply
- Explain and graphically illustrate market equilibrium, surplus and shortage

Module 4: Applications of Supply and Demand
- Analyze the economic effect of government setting price ceilings and floors
- Define, calculate, and illustrate consumer, producer, and total surplus
- Examine ways that supply and demand apply to labor and financial markets

Module 5: Elasticity
- Explain the concept of elasticity
- Explain the price elasticity of demand and price elasticity of supply, and compute both using the midpoint method
- Explain and calculate other elasticities using common economic variables
- Explain the relationship between a firm’s price elasticity of demand and total revenue

Module 6: Utility
- Describe the concept of utility and explain how consumers spend in order to maximize utility
- Explain how consumer utility changes when income or prices change
- Find consumer equilibrium using indifference curves and a budget constraint
- Describe the behavioral economics approach to understanding decision making

Module 7: Production and Costs

- Explain production and the production function
- Calculate, graph, and understand production costs in the short run
- Examine production choices in the long run

Module 8: Perfect Competition

- Describe the characteristics of perfect competition and calculate costs, including fixed, variable, average, marginal, and total costs
- Analyze a firm’s profit margin
- Describe how perfectly competitive markets adjust to long run equilibrium

Module 9: Monopoly

- Describe characteristics of monopolies
- Calculate and graph a monopoly’s costs, revenues, profit and losses
- Analyze strategies used to control monopolies

Module 10: Monopolistically Competition and Oligopoly

- Describe the characteristics of a monopolistically competitive industry
- Calculate and graph a firm’s fixed, variable, average, marginal, and total costs in monopolistic competition
- Describe characteristics of oligopolies

Module 11: Public Goods

- Define and give examples of public goods and externalities
- Define and give examples of positive and negative externalities
- Analyze the efficacy of government policies to lessen negative externalities
- Analyze how the government promotes positive externalities

Module 12: Labor Markets

- Analyze labor markets and how supply and demand interact to determine the market wage rate
- Explain how wages are determined when employers or employees hold labor market power
- Analyze the economic implications of discrimination and immigration policies

Module 13: Income Distribution

- Explain poverty and the poverty trap
- Analyze and measure economic inequality

Module 14: Globalization and Trade

- Define and calculate comparative advantage, and understand how countries choose which goods and services to trade internationally
Explain how barriers to trade (like tariffs, quotas and non-tariff barriers) affect businesses, consumers and workers in the economy
Differentiate between alternative international trade regimes and how they impact global trade

Module 15: Exchange Rates and International Finance

Define currency exchange rates and explain how they influence trade balances
Analyze how supply and demand affects foreign currencies and exchange rates
Explain how the balance of trade (surplus or deficit) affects the domestic economy
Why study economics?

Many students find the prospect of taking an economics course daunting (or maybe just dull). At the heart of this worry is perhaps just a misperception of what economics is about. It’s not rocket science, it’s not a collection of boring facts, and it’s not the study of money or the stock market. Economics is really just a set of interesting questions organized around a simple fact: there aren’t enough resources (money, land, time, etc.) to go around or satisfy all our needs and desires. Economists call this condition scarcity. It affects individuals, nations, and the entire human species—no one ever has enough of the things they want. On some level, everyone has to grapple with scarcity, and economists are interested in understanding how people do that.

If you understand how people behave in the face of scarcity—and learn to think like an economist—economics can be an amazingly powerful tool. You can predict the behavior of individual economic agents, such as consumers or businesses—what economists call the micro level. You can predict the behavior of an economy (or economies) as a whole—what economists call the macro level. You can have a better understanding of the choices—and consequences—in your own life.
Consider the following example:

Imagine that you’re about to catch a flight to Italy. You’ve saved and saved to pay for this trip, and you’re thrilled to finally be going. You’re on top of the world, until . . .

You get to the airport and have to go through security. The line is terrible. What choice do you have? You can’t board the plane without passing security. As you wait, you notice a different aisle for “special” passengers who fly more often. They aren’t waiting at all. In fact, if more than three of them collect in the line and have to wait for more than a minute, they act very annoyed—shifting their weight, rolling their eyes, checking their phones, and so on. Oh, brother! You’ve been waiting so much longer! How is that fair?

Finally you make it through security and reach your gate. Sadly, you are in zone 5, which boards last. You have to struggle down the aisle—past rows of seats with more leg room—to a center seat. Worse, those who boarded before you have filled all the overhead bins. A flight attendant seems irritated that you have a large carry-on bag that won’t fit under your seat. He takes your bag off the plane and tells you to pick it up at baggage claim after the flight.

You tuck in your elbows and squeeze into your seat thinking, “This isn’t fair.”

Not sure what all this has to do with scarcity? You need to study economics!

---

INTRODUCTION TO ECONOMICS

What you’ll learn to do: explain what economics is and explain why it is important

Most people think of money when they think of economics. While that is certainly one aspect of it, economics is about a lot more than money. Really, it is a study about decision-making and choices, and how scarcity and competition lead people to behave.
Watch It

Watch this video for a brief overview of economics. You'll often encounter videos like this, enclosed inside of “Watch It” boxes, in this course. These are highly recommended and serve as wonderful tools to review and enhance your understanding of the concepts presented. Even if you feel like you understand the material in the reading, it is worthwhile to watch the videos.
Watch this video online: https://youtu.be/g9uUIUqhrSQ

In order to understand economics, it's important to master a set of key definitions and understand how they interconnect. These concepts will be used many times throughout the course. At the most basic level:

- **Scarcity** means that there are never enough resources to satisfy all human wants
- **Economics** is the study of the trade-offs and choices that we make, given the fact of scarcity
- **Opportunity cost** is what we give up when we choose one thing over another

In this section, we will spend more time with these definitions, and understand how they’re used in the context of this discipline.

---

**UNDERSTANDING ECONOMICS AND SCARCITY**

**Learning Objectives**

- Describe scarcity and explain its economic impact
- Describe factors of production
Figure 1. Food, like the wheat shown here, is a scarce good because it exists in limited supply.

**Scarcity**

The resources that we value—time, money, labor, tools, land, and raw materials—exist in limited supply. There are simply never enough resources to meet all our needs and desires. This condition is known as scarcity.

At any moment in time, there is a finite amount of resources available. Even when the number of resources is very large, it's limited. For example, according to the U.S. Bureau of Labor Statistics, in 2016, the labor force in the United States contained more than 158 million workers—that's a lot, but it's not infinite. Similarly, the total area of the United States is 3,794,101 square miles—an impressive amount of acreage, but not endless. Because these resources are limited, so are the numbers of goods and services we can produce with them. Combine this with the fact that human wants seem to be virtually infinite, and you can see why scarcity is a problem.

**Try It**

Throughout the course, you will find these “Try It” boxes with questions to help you check your understanding and apply the concepts from the reading. Choose an answer, then select “check answer” to get feedback about how you did.

**Economics**

When faced with limited resources, we have to make choices. Again, economics is the study of how humans make choices under conditions of scarcity. These decisions can be made by individuals, families, businesses, or societies.

Let's consider a few decisions that we make based on limited resources. Take the following:

1. **What classes are you taking this term?**

Are you the lucky student who is taking every class you wanted with your first-choice professor during the perfect time and at the ideal location? The odds are that you have probably had to make trade-offs on account of scarcity. There is a limited number of time slots each day for classes and only so many faculty available to teach them. Every faculty member can't be assigned to every time slot. Only one class can be assigned to each classroom at a given time. This means that each student has to make trade-offs between the time slot, the instructor, and the class location.

2. **Where do you live?**

Think for a moment, if you had all the money in the world, where would you live? It's probably not where you're living today. You have probably made a housing decision based on scarcity. What location did you pick? Given limited time, you may have chosen to live close to work or school. Given the demand for housing, some locations are more expensive than others, though, and you may have chosen to spend more money for a convenient location or to spend
less money for a place that leaves you spending more time on transportation. There is a limited amount of housing in any location, so you are forced to choose from what's available at any time. Housing decisions always have to take into account what someone can afford. Individuals making decisions about where to live must deal with limitations of financial resources, available housing options, time, and often other restrictions created by builders, landlords, city planners, and government regulations.

Watch It: Scarcity and Choice

Throughout this course you'll encounter a series of short videos that explain complex economic concepts in very simple terms. Take the time to watch them! They'll help you master the basics and understand the readings (which tend to cover the same information in more depth).

Watch this video online: https://youtu.be/yoVc_S_gd_0

As you watch the video, consider the following key points:

1. Economics is the study of how humans make choices under conditions of scarcity.
2. Scarcity exists when human wants for goods and services exceed the available supply.
3. People make decisions in their own self-interest, weighing benefits and costs.

The Problem of Scarcity

Every society, at every level, must make choices about how to use its resources. Families must decide whether to spend their money on a new car or a fancy vacation. Towns must choose whether to put more of the budget into police and fire protection or into the school system. Nations must decide whether to devote more funds to national defense or to protecting the environment. In most cases, there just isn't enough money in the budget to do everything.

Economics helps us understand the decisions that individuals, families, businesses, or societies make, given the fact that there are never enough resources to address all needs and desires.

Economic Goods and Free Goods

Most goods (and services) are economic goods, i.e. they are scarce. Scarce goods are those for which the supply would be greater than the demand if their price were zero. Because of this shortage, economic goods have a positive price in the market. That is, consumers have to pay to get them.

What is an example of a good which is not scarce? Water in the ocean? Sand in the desert? Any good whose supply is greater than the demand if their price were zero is called a free good, since consumers can obtain all they want at no charge. We used to consider air a free good, but increasingly clean air is scarce.

Productive Resources

Having established that resources are limited, let's take a closer look at what we mean when we talk about resources. There are four productive resources (resources have to be able to produce something), also called factors of production:
- **Land**: any natural resource, including actual land, but also trees, plants, livestock, wind, sun, water, etc.
- **Economic capital**: anything that’s manufactured in order to be used in the production of goods and services. Note the distinction between financial capital (which is not productive) and economic capital (which is). While money isn’t directly productive, the tools and machinery that it buys can be.
- **Labor**: any human service—physical or intellectual. Also referred to as *human capital*.
- **Entrepreneurship**: the ability of someone (an entrepreneur) to recognize a profit opportunity, organize the other factors of production, and accept risk.

### Try It
Visit this page in your course online to view this presentation.

### Watch It
Productive resources and factors of production are explained again in more detail in the following video:
Watch this video online: [https://youtu.be/0PgP0dXAGAE](https://youtu.be/0PgP0dXAGAE)

### Glossary
- **economic goods**: goods or services a consumer must pay to obtain; also called scarce goods
- **free goods**: goods or services that a consumer can obtain for free because they are abundant relative to the demand
- **productive resources**: the inputs used in the production of goods and services to make a profit: land, economic capital, labor, and entrepreneurship; also called “factors of production”

---

**THE CONCEPT OF OPPORTUNITY COST**

### Learning Objectives
- Describe opportunity cost and its importance in decision-making
The Idea of Opportunity Cost

Since resources are limited, every time you make a choice about how to use them, you are also choosing to forego other options. Economists use the term opportunity cost to indicate what must be given up to obtain something that's desired. A fundamental principle of economics is that every choice has an opportunity cost. If you sleep through your economics class (not recommended, by the way), the opportunity cost is the learning you miss. If you spend your income on video games, you cannot spend it on movies. If you choose to marry one person, you give up the opportunity to marry anyone else. In short, opportunity cost is all around us.

The idea behind opportunity cost is that the cost of one item is the lost opportunity to do or consume something else; in short, opportunity cost is the value of the next best alternative.

Since people must choose, they inevitably face trade-offs in which they have to give up things they desire to get other things they desire more.

Opportunity Cost and Individual Decisions

In some cases, recognizing the opportunity cost can alter personal behavior. Imagine, for example, that you spend $8 on lunch every day at work. You may know perfectly well that bringing a lunch from home would cost only $3 a day, so the opportunity cost of buying lunch at the restaurant is $5 each day (that is, the $8 that buying lunch costs minus the $3 your lunch from home would cost). Five dollars each day does not seem to be that much. However, if you project what that adds up to in a year—250 workdays a year × $5 per day equals $1,250—it's the cost, perhaps, of a decent vacation. If the opportunity cost were described as “a nice vacation” instead of “$5 a day,” you might make different choices.

Try It

Visit this page in your course online to view this presentation.

Opportunity Cost and Societal Decisions

Opportunity cost also comes into play with societal decisions. Universal health care would be nice, but the opportunity cost of such a decision would be less housing, environmental protection, or national defense. These trade-offs also arise with government policies. For example, after the terrorist plane hijackings on September 11, 2001, many proposals, such as the following, were made to improve air travel safety:
- The federal government could provide armed “sky marshals” who would travel inconspicuously with the rest of the passengers. The cost of having a sky marshal on every flight would be roughly $3 billion per year.
- Retrofitting all U.S. planes with reinforced cockpit doors to make it harder for terrorists to take over the plane would have a price tag of $450 million.
- Buying more sophisticated security equipment for airports, like three-dimensional baggage scanners and cameras linked to face-recognition software, would cost another $2 billion.

However, the single biggest cost of greater airline security doesn’t involve money. It’s the opportunity cost of additional waiting time at the airport. According to the United States Department of Transportation, more than 800 million passengers took plane trips in the United States in 2012. Since the 9/11 hijackings, security screening has become more intensive, and consequently, the procedure takes longer than in the past. Say that, on average, each air passenger spends an extra 30 minutes in the airport per trip. Economists commonly place a value on time to convert an opportunity cost in time into a monetary figure. Because many air travelers are relatively highly paid businesspeople, conservative estimates set the average “price of time” for air travelers at $20 per hour. Accordingly, the opportunity cost of delays in airports could be as much as 800 million (passengers) × 0.5 hours × $20/hour—or, $8 billion per year. Clearly, the opportunity costs of waiting time can be just as substantial as costs involving direct spending.

**Figure 2. Time and Money.** Lost time can be a significant component of opportunity cost.

**Try It**

Visit this page in your course online to view this presentation.

**Watch It: Opportunity Cost**

Watch this video to see some more examples and a deeper explanation of opportunity cost.

Watch this video online: [https://youtu.be/PSU-_n81QT0](https://youtu.be/PSU-_n81QT0)

**Glossary**

**opportunity cost:** the value of the next best alternative

**License & Attributions**

CC licensed content, Original

- Revision and adaptation. Provided by: Lumen Learning. License: CC BY: Attribution

CC licensed content, Shared previously

- How Individuals Make Choices Based on Their Budget Constraint. Authored by: OpenStax College. Located at: [https://cnx.org/contents/vEmOH-_p84.4448ejHQax@9/How-Individuals-Make-Choices-B](https://cnx.org/contents/vEmOH-_p84.4448ejHQax@9/How-Individuals-Make-Choices-B). License: CC BY: Attribution. License Terms: Download for free at [http://cnx.org/contents/bc49bf1f-efe9-43a0-8dea-d335ef9ad09a82@4.44](http://cnx.org/contents/bc49bf1f-efe9-43a0-8dea-d335ef9ad09a82@4.44)

**LABOR, MARKETS, AND TRADE**
Learning Objectives

- Explain how specialization and the division of labor leads to trade and markets

The Division and Specialization of Labor

![Factory. An assembly line.](image)

We have learned that there aren’t enough resources to fulfill all of our wants and this reality forces us to make choices that have opportunity costs. How do we get the most we can from the resources we have? Over time, markets and trade have come into existence and have become highly efficient mechanisms for optimizing our use of resources and bringing us the most and best combination of goods and services.

Think back to pioneer days, when the average person knew how to do so much more on his or her own than someone today—everything from shoeing a horse to growing, hunting, and preserving food to building a house and repairing equipment. Most of us don’t know how to do all—or any—of those things. It’s not because we’re not capable of learning them. It’s because we don’t have to. The reason for this is something called the “division and specialization of labor,” a production innovation first put forth by Adam Smith.

The formal study of economics began when Adam Smith (1723–1790) published his famous book, *The Wealth of Nations*, in 1776. Many authors had written about economics in the centuries before Smith, but he was the first to address the subject in a comprehensive way.

In the first chapter of the book, Smith introduces the idea of the division of labor, which means that the way a good or service is produced is divided into a number of tasks that are performed by different workers, instead of all the tasks being performed by the same person. To illustrate the division of labor, Smith counted how many tasks were involved in making a pin: drawing out a piece of wire, cutting it to the right length, straightening it, putting a head on one end and a point on the other, packaging pins for sale, and so on. Smith counted eighteen distinct tasks that were typically performed by different people—all for a pin!

Modern companies divide tasks, too. Even a relatively simple business like a restaurant divides up the task of serving meals into a range of jobs: top chef, sous chefs, less-skilled kitchen help, host/hostess, waiters/waitresses, janitors, a business manager to handle accounts and paychecks, etc. A complex business like a large manufacturing factory or a hospital can have hundreds of job classifications.
Why the Division of Labor Increases Production

When the tasks involved with producing a good or service are divided and subdivided, workers and businesses can produce a greater quantity of those goods or services. In his study of pin factories, Smith observed that one worker alone might make twenty pins in a day, but that a small business of ten workers (some of whom would need to do two or three of the eighteen tasks involved in pin making), could make forty-eight thousand pins in a day. How can a group of workers, each specializing in certain tasks, produce so much more than the same number of workers who try to produce the entire good or service by themselves? Smith offered three reasons. First, specialization in a particular small job allows workers to focus on the parts of the production process in which they have an advantage. People have different skills, talents, and interests, so they will be better at some jobs than at others. The particular advantages may be based on educational choices, which are shaped, in turn, by interests and talents. Only those with medical training qualify to become doctors, for instance. For some goods, specialization will be affected by geography—it’s easier to be a wheat farmer in North Dakota than in Florida, but easier to run a tourist hotel in Florida than in North Dakota. If you live in or near a big city, it’s easier to attract enough customers to operate a successful dry-cleaning business or movie theater than if you live in a sparsely populated rural area. Whatever the reason, if people specialize in the production of what they do best, they will be more productive than if they produce a combination of things, some of which they are good at and some of which they are not.

Second, workers who specialize in certain tasks often learn to produce more quickly and with higher quality. This pattern holds true for many workers, including assembly-line laborers who build cars, stylists who cut hair, and doctors who perform heart surgery. In fact, specialized workers often know their jobs well enough to suggest innovative ways to do their work faster and better. A similar pattern often operates within businesses. In many cases, a business that focuses on one or a few products is more successful than firms that try to make a wide range of products.

Third, specialization allows businesses to take advantage of economies of scale, which means that, for many goods, as the level of production increases, the average cost of producing each individual unit declines. For example, if a factory produces only one hundred cars per year, each car will be quite expensive to make on average. However, if a factory produces fifty thousand cars each year, then it can set up an assembly line with huge machines and workers performing specialized tasks, and the average cost of production per car will drop. Economies of scale implies that production is becoming more efficient as the scale of production rises.

The ultimate result of workers who can focus on their preferences and talents, learn to do their specialized jobs better, and work in larger organizations is that society as a whole can produce and consume far more than if each person tried to produce all of their own goods and services. The division and specialization of labor has been a force against the problem of scarcity.

Trade and Markets

Specialization only makes sense, though, if workers (and other economic agents such as businesses and nations) can use their income to purchase the other goods and services they need. In short, specialization requires trade. You do not have to know anything about electronics or sound systems to play music—you just buy an iPod or MP3 player, download the music, and listen. You don’t have to know anything about textiles or the construction of sewing machines if you need a jacket—you just buy the jacket and wear it. Instead of trying to acquire all the knowledge and skills involved in producing all of the goods and services that you wish to consume, the market allows you to learn a specialized set of skills and then use the pay you receive to buy the goods and services you need or want. This is how our modern society has evolved into a strong economy.
MICROECONOMICS AND MACROECONOMICS

Learning Objectives

- Distinguish between macroeconomics and microeconomics

Micro vs. Macro

It should be clear by now that economics covers a lot of ground. That ground can be divided into two parts: microeconomics focuses on the actions of individual agents within the economy, like households, workers, and businesses; macroeconomics looks at the economy as a whole. It focuses on broad issues such as growth, unemployment, inflation, and trade balance. Microeconomics and macroeconomics are not separate subjects but are, rather, complementary perspectives on the overall subject of the economy.

To understand why both microeconomic and macroeconomic perspectives are useful, consider the problem of studying a biological ecosystem like a lake. One person who sets out to study the lake might focus on specific topics: certain kinds of algae or plant life; the characteristics of particular fish or snails; or the trees surrounding the lake. Another person might take an overall view and instead consider the entire ecosystem of the lake from top to bottom: what eats what, how the system remains in balance, and what environmental stresses affect this balance. Both approaches are useful, and both researchers study the same lake, but the viewpoints are different. In a similar way, both microeconomics and macroeconomics study the same economy, but each has a different starting point, perspective, and focus.

Whether you are looking at lakes or economics, the micro and the macro insights should illuminate each other. In studying a lake, the “micro” insights about particular plants and animals help us to understand the overall food chain,
while the “macro” insights about the overall food chain help to explain the environment in which individual plants and animals live.

In economics, the micro decisions of individual businesses are influenced by the health of the macroeconomy—for example, firms will be more likely to hire workers if the overall economy is growing. In turn, the performance of the macroeconomy ultimately depends on the microeconomic decisions made by individual households and businesses.

Microeconomics

What determines how households and individuals spend their budgets? What combination of goods and services will best fit their needs and wants, given the budget they have to spend? How do people decide whether to work, and if so, whether to work full time or part time? How do people decide how much to save for the future, or whether they should borrow to spend beyond their current means?

What determines the products, and how many of each, a firm will produce and sell? What determines what prices a firm will charge? What determines how a firm will produce its products? What determines how many workers it will hire? How will a firm finance its business? When will a firm decide to expand, downsize, or even close? In the microeconomic part of this text, we will learn about the theory of consumer behavior and the theory of the firm.

Macroeconomics

What determines the level of economic activity in a society or nation?—that is, how many goods and services does it actually produce? What determines how many jobs are available in an economy? What determines a nation’s standard of living? What causes the economy to speed up or slow down? What causes firms to hire more workers or lay them off? Finally, what causes the economy to grow over the long term?

An economy’s macroeconomic health can be assessed by a number of standards or goals. The most important macroeconomic goals are the following:

- Growth in the standard of living
- Low unemployment
- Low inflation

Macroeconomic policy pursues these goals through monetary policy and fiscal policy:

- **Monetary policy**, which involves policies that affect bank lending, interest rates, and financial capital markets, is conducted by a nation’s central bank. For the United States, this is the Federal Reserve.
- **Fiscal policy**, which involves government spending and taxes, is determined by a nation’s legislative body. For the United States, this is the Congress and the executive branch, which establishes the federal budget.

To keep the differences between these policies straight, remember that the term monetary relates to money, and the term fiscal relates to government revenue or taxes.

These are the main tools the government has to work with. Americans tend to expect that government can fix whatever economic problems we encounter, but to what extent is that expectation realistic? These are just some of the issues that will be explored later in this course.

**Try It**

Visit this page in your course online to view this presentation.

**Watch It**
The differences between microeconomics and macroeconomics as well as their respective focal points are explained again in the following video:

Watch this video online: https://youtu.be/w8tUIq7Blsg

**Glossary**

**fiscal policy**: economic policies that involve government spending and taxes

**macroeconomics**: the branch of economics that focuses on broad issues such as growth, unemployment, inflation, and trade balance

**microeconomics**: the branch of economics that focuses on actions of particular agents within the economy, like households, workers, and businesses

**monetary policy**: policy that involves altering the level of interest rates, the availability of credit in the economy, and the extent of borrowing

**INTRODUCTION TO MATH IN ECONOMICS**

What you’ll learn to do: use mathematics in common economic applications
Earlier we defined economics as the study of how people choose to use scarce resources to best satisfy their unlimited wants. Economists try to analyze these choices both at the individual level (what wage does Alissa require in order to take that job?) and broader societal level (what is the impact of minimum wage on the unemployment rate?). The principal tools economists use to do this are models.

Economists use math as a tool for manipulating and exploring economic models. Sometimes it makes sense to express economic ideas in words; other times, math does a better job. Economics is not math, but rather math is a tool for presenting and manipulating/exploring/using economic models. Many economic models use math to explain cause and effect. Don’t worry, though, we’re going to cover all the math you need to solve the problems in this course.

This section provides a quick review of some basic math (so you can avoid common errors) and then introduces the mathematical concepts you’ll need throughout the course. Do**n’t forget: return to this section later on if you encounter math that you can’t follow.**

Some students, when they find out that economics involves math, fear that the math will trip them up and prevent their success in the course—“I’m not a math person!” they think. If you share these thoughts, it may surprise you to know there’s scientific research showing that when you practice new ways of thinking, your brain physically changes and adapts. Essentially, there’s no such thing as “a math person” (or an “economics person”). You don’t need a special talent or aptitude. It’s mainly a matter of practice, hard work, and training your brain. The more you challenge your mind to learn, the more your brain cells connect to one another and the stronger those connections become.

So, how do you actually develop your brain and succeed in this kind of course? We asked some former students to tell us their advice for success. They said it was most important to

- Do the practice problems in the course.
- Ask questions.
- Study your mistakes.
- Explain the ideas to yourself in different ways until they are clear.

You will need to learn new things to pass this class, and you should expect it to feel hard as you wrestle with unfamiliar ideas and new ways of thinking. Don’t give up, though! The feeling of struggling is a normal part of how the brain gets stronger when it learns things.
ECONOMIC MODELS

Learning Objectives

- Explain the characteristics and purpose of economic models

An **economic model** is a simplified version of reality that allows us to observe, understand, and make predictions about economic behavior. The purpose of a model is to take a complex, real-world situation and pare it down to the essentials. If designed well, a model can give the analyst a better understanding of the situation and any related problems.

A good model is simple enough to be understood while complex enough to capture key information. Sometimes economists use the term *theory* instead of *model*. Strictly speaking, a theory is a more abstract representation, while a model is a more applied or empirical representation. Often, models are used to test theories. In this course, however, we will use the terms interchangeably.

Watch It

Watch this video to get a better grasp on economic models and why they are useful to economists in making predictions about behavior.

Watch this video online: [https://youtu.be/IYHy2XLa4u0](https://youtu.be/IYHy2XLa4u0)

Economic Models and Math

Economists use models as the primary tool for explaining or making predictions about economic issues and problems. For example, an economist might try to explain what caused the Great Recession in 2008, or she might try to predict how a personal income tax cut would affect automobile purchases.

Economic models can be represented using words or using mathematics. All of the important concepts in this course can be explained without math. That said, math is a tool that can be used to explore economic concepts in very helpful ways. You know the saying "A picture is worth a thousand words"? The same applies to graphs: they’re a very effective means of conveying information visually—without a thousand words. In addition to being a “picture,” a graph is also a math-based model.

The use of algebra is a specific way that economics express and explore economic models. Where graphs require you to "eyeball" a model, algebra can give you more precise answers to questions. For example, if a business puts their product on sale for 10% off the regular price, how much more will consumers buy? Similarly, using the algebraic formula for a line allows economists to find precise points on a graphs that help in interpreting how much of a good should be sold, or at what price.

Why would an economist use math when there are other ways of representing models, such as with text or narrative? Why would you use your fist to bang a nail, if you had a hammer? Math has certain advantages over text. It disciplines our thinking by making us specify exactly what we mean. You can get away with fuzzy thinking and vague approximations in your own mind, but not when you’re reducing a model to algebraic equations. At the same time, math has certain disadvantages. Mathematical models lack the nuances that can be found in narrative models. The point is that math is one tool, but it’s not the only tool or even always the best tool economists can use to work with economic models.

Try It
Examples of Models

An architect who is designing a major office building will probably build a physical model that sits on a tabletop to show how the entire city block will look after the new building is constructed. Companies often build models of their new products that are rougher and less finished than the final product but can still demonstrate how the new product will work and look. Such models help people visualize a product (or a building) in a more complete, concrete way than they could without them.

Similarly, economic models offer a way to get a complete view or picture of an economic situation and understand how economic factors fit together.

A good model to start with in economics is the circular flow diagram (Figure 2, below). Such a diagram indicates that the economy consists of two groups, households and firms, which interact in two markets: the goods-and-services market (also called the product market), in which firms sell and households buy, and the labor market, in which households sell labor to business firms or other employees.

Of course, in the real world, there are many different markets for goods and services and markets for many different types of labor. The circular flow diagram simplifies these distinctions in order to make the picture easier to grasp. In the diagram, firms produce goods and services, which they sell to households in return for payments. The outer ring represents the two sides of the product market (which provides goods and services), in which households demand and firms supply. In addition, households (as workers) sell their labor to firms in return for wages, salaries, and benefits. This is shown in the inner circle, which represents the two sides of the labor market, in which households supply and firms demand. This version of the circular flow model is stripped down to the essentials, but it has enough features to explain how the product and labor markets work in the economy.

We could easily add details to this basic model if we wanted to introduce more real-world elements, like financial markets, governments, or interactions with the rest of the world (imports and exports). Economists reach for theories in much the same way as a carpenter might grab a tool. When economists identify an economic issue or problem, they sift through the available theories to see if they can find one that fits. Then they use the theory to give them insights about the issue or problem. In economics, theories are expressed in models as diagrams, graphs, or even as mathematical equations. Counter to what you might expect, economists don’t figure out the solution to a problem and then draw the graph. Instead, they use the graph to help them discover the answer. In this way, these graphs serve as models to make inferences about behavior.
At the introductory level, you can sometimes figure out the right answer without using a model, but if you keep studying economics, before too long you’ll encounter issues and problems whose solution will require graphs. Both micro and macroeconomics are explained in terms of theories and models. The most well-known theories are probably those of supply and demand, but you will learn about several others.

Try It

Visit this page in your course online to view this presentation.

Glossary

circular flow diagram: a diagram indicating that the economy consists of households and firms interacting in a goods-and-services market and a labor market

goods and services market: a market in which firms are sellers of what they produce and households are buyers

economic model: is a simplified version of reality that allows us to observe, understand, and make predictions about economic behavior

labor market: the market in which households sell their labor as workers to businesses or other employers

PURPOSE OF FUNCTIONS

Learning Objectives

- Explain how equations and functions are used to describe relationships (such as cause and effect)

Often, economic models are expressed in terms of mathematical functions. What’s a function? Basically, a function describes a relationship involving one or more variables. Sometimes the relationship is a definition. For example (using words), Joan of Arc is a professor. This could be expressed as Joan of Arc = professor. Or, food = cherries, cheese, and chocolate means that cherries, cheese, and chocolate are food.

In economics, functions frequently describe cause and effect. The variable on the left-hand side is what is being explained (“the effect”). On the right-hand side is what’s doing the explaining (“the causes”). Functions are also useful for making predictions. For example, think about your grade in this course. We might be able to predict how well you will do in this course by considering how well you’ve done in other courses, by how much you attend class or participate in the online activities, and by how many hours you study.
Not all of those things will have equal impact on your grade. Let's assume that your study time is most important and will have twice as much impact as the other factors. We are trying to describe 100 percent of the impact, so study time will explain 50 percent, attendance and participation will explain 25 percent, and your prior class grades will describe 25 percent. Together, this adds up to 100 percent.

Now, let's turn that into a function. Your grade in the course can be represented as the following:

\[
\text{Grade} = (0.50 \times \text{hours spent studying}) + (0.25 \times \text{class attendance}) + (0.25 \times \text{prior GPA})
\]

This equation states that your grade depends on three things: the number of hours you spend studying, your class attendance, and your prior course grades represented as your grade-point average (GPA). It also says that study time is twice as important (0.50) as either class attendance (0.25) or prior GPA score (0.25). If this relationship is true, how could you raise your grade in this course? By not skipping class and studying more. Note that you cannot do anything about your prior GPA, since that is calculated from courses you've already taken and grades you've already received.

Try It

Visit this page in your course online to view this presentation.

Economic models tend to express relationships using economic variables, such as:

\[
\text{Budget} = \text{money spent on econ books} + \text{money spent on music}
\]

(assuming that the only things you buy are economics books and music). Often, there is some assumption that has to be explained in order to identify where the model has been simplified.

As you can see, in economic models the math isn't difficult. It's used to help describe and explain the relationships between variables.

Glossary

function: a relationship or expression involving one or more variables
Let's quickly review some math concepts that will help you avoid simple errors in your work.

Order of Operations

Remember, when you solve an equation it’s important to do each operation in the following order:

1. Simplify inside parentheses and brackets.
2. Simplify the exponent.
3. Multiply and divide from left to right.
4. Add and subtract from left to right.

In this course you will not use exponents, but you will need to remember the order of the other steps. So, in solving the following equation, you multiply first, then add:

\[ y = 9 + 3 \times 10 \]
\[ y = 9 + 30 \]
\[ y = 39 \]

Try It

These next two questions allow you to get as much practice as you need, as you can click the link at the top of each question ("Try another version of this question") to get a new question. Practice until you feel comfortable doing the questions and then move on.

Visit this page in your course online to practice before taking the quiz.

Visit this page in your course online to practice before taking the quiz.

Try It

These next two problems are a little harder. See if you can do them. Remember to apply the order of operations: simplify inside parentheses first; then multiply and divide left to right; finally add and subtract left to right.

Visit this page in your course online to view this presentation.

Visit this page in your course online to view this presentation.

Lines

In this course the most common equation you will see is \( y = b + mx \). This is the equation for a line. We will revisit this equation later in this module when we review graphs. For now, let’s practice solving this common equation using different variables.

Understanding Variables

To a mathematician or an economist, a **variable** is the name given to a quantity that can assume a range of values. In other words, the value of a variable can change or vary. In an equation it’s represented by a letter or a symbol. Because economic models often consider cause and effect, variables are important. You will often be asked to consider a range of options that result from different variables. Below is a very simple example:
In order to understand the range of options, we might start with 0. What does y equal if \( x = 0 \)?

\[
y = 9 + 3x \\
y = 9 + 3(0) \\
y = 9 + 0 \\
y = 9
\]

Now, let's look at the same formula with different information. What does y equal if \( x = 5 \)?

\[
y = 9 + 3x \\
y = 9 + 3(5) \\
y = 9 + 15 \\
y = 24
\]

**Working with Variables**

Remember that when you’re trying to solve an equation with one or more variables, you need to isolate the variable. Let’s walk through a simple example using the same equation from above. What if we want to solve the equation in a case where \( y = 24 \)?

\[
y = 9 + 3x \\
24 = 9 + 3x
\]

First, subtract the same number from each side of the equation to simplify the equation without changing the fact that it’s an equality. In this case, we want to subtract the number that will enable us to isolate \( x \) (x is on one side of the equal sign all by itself). We can do that by subtracting 9 from each side.

\[
24 = 9 + 3x \\
-9 = -9 \\
15 = 3x
\]

Now we can further simplify the equation by dividing both sides by 3.

\[
\frac{15}{3} = \frac{3x}{3} \\
5 = x
\]

Let’s practice solving for \( x \) one more time. What does \( x \) equal if \( y = 12 \)?

\[
12 = 9 + 3x \\
-9 = -9 \\
3 = 3x \\
\frac{3}{3} = \frac{3x}{3}
\]
INTRODUCTION TO GRAPHS IN ECONOMICS

What you’ll learn to do: use graphs in common economic applications

In this course, the most common way you will encounter economic models is in graphical form.

A graph is a visual representation of numerical information. Graphs condense detailed numerical information to make it easier to see patterns (such as “trends”) among data. For example, which countries have larger or smaller populations? A careful reader could examine a long list of numbers representing the populations of many countries, but with more than two hundred nations in the world, searching through such a list would take concentration and time. Putting these same numbers on a graph, listing them from highest to lowest, would reveal population patterns much more readily.
Economists use graphs not only as a compact and readable presentation of data, but also for visually representing relationships and connections—in other words, they function as models. As such, they can be used to answer questions. For example: How do increasing interest rates affect home sales? Graphing the results can help illuminate the answers.

This section provides an overview of graphing—just to make sure you're up to speed on the basics. It's important to feel comfortable with the way graphs work before using them to understand new concepts.

---

**CREATING AND INTERPRETING GRAPHS**

**Learning Objectives**

- Explain how to construct a simple graph that shows the relationship between two variables

It's important to know the terminology of graphs in order to understand and manipulate them. Let's begin with a visual representation of the terms (shown in Figure 1), and then we can discuss each one in greater detail.

---

*Figure 1. Graph Terminology.*
Throughout this course we will refer to the horizontal line at the base of the graph as the \textbf{x-axis}. We will refer to the vertical line on the left hand side of the graph as the \textbf{y-axis}. This is the standard convention for graphs. In economics, we commonly use graphs with price (p) represented on the x-axis, and quantity (q) represented on the y-axis.

An \textbf{intercept} is where a line on a graph crosses (“intercepts”) the x-axis or the y-axis. Mathematically, the x-intercept is the value of x when y = 0. Similarly, the y-intercept is the value of y when x = 0. You can see the x-intercepts and y-intercepts on the graph above.

The point where two lines on a graph cross is called an \textbf{intersection point}.

The other important term to know is \textbf{slope}. The slope tells us how steep a line on a graph is as we move from one point on the line to another point on the line. Technically, \textbf{slope} is the change in the vertical axis divided by the change in the horizontal axis. The formula for calculating the slope is often referred to as the “rise over the run”—again, the change in the distance on the y-axis (rise) divided by the change in the x-axis (run).

Now that you know the “parts” of a graph, let’s turn to the equation for a line:

\[ y = mx + b \]

In any equation for a line, \( m \) is the slope and \( b \) is the y-intercept.

Let’s use the same equation we used earlier, in the section on solving algebraic equations, \( y = 9 + 3x \), which can also be written as:

\[ y = 3x + 9 \]

In this equation for a line, the \( b \) term is 9 and the \( m \) term is 3. The table below shows the values of \( x \) and \( y \) for this equation. To construct the table, just plug in a series of different values for \( x \), and then calculate the resulting values for \( y \).

<table>
<thead>
<tr>
<th>( x )</th>
<th>( y )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td>1</td>
<td>12</td>
</tr>
<tr>
<td>2</td>
<td>15</td>
</tr>
<tr>
<td>3</td>
<td>18</td>
</tr>
<tr>
<td>4</td>
<td>21</td>
</tr>
<tr>
<td>5</td>
<td>24</td>
</tr>
<tr>
<td>6</td>
<td>27</td>
</tr>
</tbody>
</table>

Next we can place each of these points on a graph. We can start with 0 on the \( x \)-axis and plot a point at 9 on the \( y \)-axis. We can do the same with the other pairs of values and draw a line through all the points, as on the graph in Figure 2, below.
This example illustrates how the \( b \) and \( m \) terms in an equation for a straight line determine the position of the line on a graph. As noted above, the \( b \) term is the \( y \)-intercept. The reason is that if \( x = 0 \), the \( b \) term will reveal where the line intercepts, or crosses, the \( y \)-axis. In this example, the line hits the vertical axis at 9. The \( m \) term in the equation for the line is the slope. Remember that slope is defined as rise over run; the slope of a line from one point to another is the change in the vertical axis divided by the change in the horizontal axis. In this example, each time the \( x \) term increases by 1 (the run), the \( y \) term rises by 3. Thus, the slope of this line is therefore \( 3/1 = 3 \). Specifying a \( y \)-intercept and a slope—that is, specifying \( b \) and \( m \) in the equation for a line—will identify a specific line. Although it is rare for real-world data points to arrange themselves as a perfectly straight line, it often turns out that a straight line can offer a reasonable approximation of actual data.

Watch It

Watch this video to take a closer look at graphs and how variables can be represented in graph form. NOTE: Around the two-minute mark, the narrator inadvertently says “indirect,” rather than “inverse.” This is corrected later in the video.

Watch this video online: https://youtu.be/uvnHPeQrk0E

Try It

Visit this page in your course online to view this presentation.

Glossary

**intercept**: the point on a graph where a line crosses the vertical axis or horizontal axis

**slope**: the change in the vertical axis divided by the change in the horizontal axis

**variable**: a quantity that can assume a range of values

**x-axis**: the horizontal line on a graph, commonly represents quantity (q) on graphs in economics

**y-axis**: the vertical line on a graph, commonly represents price (p) on graphs in economics
INTERPRETING SLOPE

Learning Objectives

- Differentiate between a positive relationship and a negative relationship

What the Slope Means

The concept of slope is very useful in economics, because it measures the relationship between two variables. A positive slope means that two variables are positively related—that is, when $x$ increases, so does $y$, and when $x$ decreases, $y$ also decreases. Graphically, a positive slope means that as a line on the line graph moves from left to right, the line rises. We will learn in other sections that "price" and "quantity supplied" have a positive relationship; that is, firms will supply more when the price is higher.
A negative slope means that two variables are negatively related; that is, when $x$ increases, $y$ decreases, and when $x$ decreases, $y$ increases. Graphically, a negative slope means that as the line on the line graph moves from left to right, the line falls. We will learn that “price” and “quantity demanded” have a negative relationship; that is, consumers will purchase less when the price is higher.

A slope of zero means that $y$ is constant no matter the value of $x$. Graphically, the line is flat; the rise over run is zero.
The unemployment-rate graph in Figure 4, below, illustrates a common pattern of many line graphs: some segments where the slope is positive, other segments where the slope is negative, and still other segments where the slope is close to zero.

Calculating Slope

The slope of a straight line between two points can be calculated in numerical terms. To calculate slope, begin by designating one point as the “starting point” and the other point as the “end point” and then calculating the rise over run between these two points.
Answer

Start from a point on the line, such as (2, 1) and move vertically until in line with another point on the line, such as (6, 3). The rise is 2 units. It is positive as you moved up.

Next, move horizontally to the point (6, 3). Count the number of units. The run is 4 units. It is positive as you moved to the right.

Then solve using the formula:

\[
\text{Slope} = \frac{\text{rise}}{\text{run}}
\]

so

\[
\text{Slope} = \frac{2}{4} = \frac{1}{2}
\]

Try It

These next questions allow you to get as much practice as you need, as you can click the link at the top of the first question (“Try another version of these questions”) to get a new set of questions. Practice until you feel comfortable doing the questions and then move on.

Visit this page in your course online to practice before taking the quiz.

Graphs of economic relationships are not always straight lines. In this course, you will often see nonlinear (curved) lines, like Figure 6, which shows the relationship between quantity of output being produced and the cost of producing that output. As the quantity of output increases, the total cost increases at a faster rate. Table 1 shows the data behind this graph.

<table>
<thead>
<tr>
<th>Table 1: Total Cost Curve</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Quantity of Output (Q)</strong></td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>&quot;Point&quot; 4</td>
</tr>
</tbody>
</table>
In this example, the total cost of production increase at a faster rate when the quantity of output increases.

We can interpret nonlinear relationships similarly to the way we interpret linear relationships. Their slopes can be positive or negative. We can calculate the slopes similarly also, looking at the rise over the run of a segment of a curve.

As an example, consider the slope of the total cost curve, above, between points A & B. Going from point A to point B, the rise is the change in total cost (i.e. the variable on the vertical axis):

$25 - $16 = $9

Similarly, the run is the change in quantity (i.e. the variable on the horizontal axis):

$5 - 4 = 1$

Thus, the slope of a straight line between these two points would be $9/1 = 9$. In other words, as we increase the quantity of output produced by one unit, the total cost of production increases by $9.

Try It

Visit this page in your course online to view this presentation.
Suppose the slope of a line were to increase. Graphically, that means it would get steeper. Suppose the slope of a line were to decrease. Then it would get flatter. These conditions are true whether or not the slope was positive or negative to begin with. A lower positive slope means a flatter upward tilt to the curve, which you can see in Figure 6 at low levels of output. A higher positive slope means a steeper upward tilt to the curve, which you can see at higher output levels.

A negative slope that is larger in absolute value (that is, more negative) means a steeper downward tilt to the line. A slope of zero is a horizontal line. A vertical line has an infinite slope.

Suppose a line has a larger intercept. Graphically, that means it would shift out (or up) from the old origin, parallel to the old line. This is shown in Figure 7, below, as the shift from the line labeled Y to the line labeled Y₁. If a line has a smaller intercept, it would shift in (or down), parallel to the old line.

**Figure 7.** A larger y-intercept shifts the entire graph to cross the y-axis at a higher point.

**Glossary**

- **negative slope**: indicates that two variables are negatively related; when one variable increases, the other decreases, and when one variable decreases, the other increases
- **positive slope**: indicates that two variables are positively related; when one variable increases, so does the other, and when one variable decreases, the other also decreases
- **slope**: the change in the vertical axis divided by the change in the horizontal axis
- **slope of zero**: indicates that there is a constant relationship between two variables: when one variable changes, the other does not change
Three types of graphs are used in this course: line graphs, pie graphs, and bar graphs. Each is discussed below.

**Line Graphs**

The graphs we’ve discussed so far are called **line graphs**, because they show a relationship between two variables: one measured on the horizontal axis and the other measured on the vertical axis.

Sometimes it’s useful to show more than one set of data on the same axes. The data in the table, below, is displayed in Figure 1, which shows the relationship between two variables: length and median weight for American baby boys and girls during the first three years of life. (The median means that half of all babies weigh more than this and half weigh less.) The line graph measures length in inches on the horizontal axis and weight in pounds on the vertical axis. For example, point A on the figure shows that a boy who is 28 inches long will have a median weight of about 19 pounds. One line on the graph shows the length-weight relationship for boys, and the other line shows the relationship for girls. This kind of graph is widely used by health-care providers to check whether a child’s physical development is roughly on track.

<table>
<thead>
<tr>
<th>Boys from Birth to 36 Months</th>
<th>Girls from Birth to 36 Months</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Length (inches)</strong></td>
<td><strong>Weight (pounds)</strong></td>
</tr>
<tr>
<td>20.0</td>
<td>8.0</td>
</tr>
<tr>
<td>22.0</td>
<td>10.5</td>
</tr>
<tr>
<td>24.0</td>
<td>13.5</td>
</tr>
<tr>
<td>26.0</td>
<td>16.4</td>
</tr>
<tr>
<td>28.0</td>
<td>19.0</td>
</tr>
<tr>
<td>30.0</td>
<td>21.8</td>
</tr>
<tr>
<td>32.0</td>
<td>24.3</td>
</tr>
<tr>
<td>34.0</td>
<td>27.0</td>
</tr>
<tr>
<td>36.0</td>
<td>29.3</td>
</tr>
</tbody>
</table>

*Figure 1. The Length-Weight Relationship for American Boys and Girls.*
Try It

Visit this page in your course online to view this presentation.

Not all relationships in economics are linear. Sometimes they are curves. Figure 2, below, presents another example of a line graph, representing the data from the table underneath. In this case, the line graph shows how thin the air becomes when you climb a mountain. The horizontal axis of the figure shows altitude, measured in meters above sea level. The vertical axis measures the density of the air at each altitude. Air density is measured by the weight of the air in a cubic meter of space (that is, a box measuring one meter in height, width, and depth). As the graph shows, air pressure is heaviest at ground level and becomes lighter as you climb. Figure 1 shows that a cubic meter of air at an altitude of 500 meters weighs approximately one kilogram (about 2.2 pounds). However, as the altitude increases, air density decreases. A cubic meter of air at the top of Mount Everest, at about 8,828 meters, would weigh only 0.023 kilograms. The thin air at high altitudes explains why many mountain climbers need to use oxygen tanks as they reach the top of a mountain.

<table>
<thead>
<tr>
<th>Altitude (meters)</th>
<th>Air Density (kg/cubic meters)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1.200</td>
</tr>
<tr>
<td>500</td>
<td>1.093</td>
</tr>
<tr>
<td>1,000</td>
<td>0.831</td>
</tr>
<tr>
<td>1,500</td>
<td>0.678</td>
</tr>
<tr>
<td>2,000</td>
<td>0.569</td>
</tr>
<tr>
<td>2,500</td>
<td>0.484</td>
</tr>
<tr>
<td>3,000</td>
<td>0.415</td>
</tr>
<tr>
<td>3,500</td>
<td>0.357</td>
</tr>
<tr>
<td>4,000</td>
<td>0.307</td>
</tr>
<tr>
<td>4,500</td>
<td>0.231</td>
</tr>
<tr>
<td>5,000</td>
<td>0.182</td>
</tr>
<tr>
<td>5,500</td>
<td>0.142</td>
</tr>
<tr>
<td>6,000</td>
<td>0.100</td>
</tr>
<tr>
<td>6,500</td>
<td>0.085</td>
</tr>
<tr>
<td>7,000</td>
<td>0.066</td>
</tr>
<tr>
<td>7,500</td>
<td>0.051</td>
</tr>
<tr>
<td>8,000</td>
<td>0.041</td>
</tr>
</tbody>
</table>

*Figure 2. Altitude–Air-Density Relationship.*
The length-weight relationship and the altitude–air-density relationship in these two figures represent averages. If you were to collect actual data on air pressure at different altitudes, the same altitude in different geographic locations would have slightly different air density, depending on factors like how far you were from the equator, local weather conditions, and the humidity in the air. Similarly, in measuring the height and weight of children for the previous line graph, children of a particular height would have a range of different weights, some above average and some below. In the real world, this sort of variation in data is common. The task of a researcher is to organize that data in a way that helps to understand typical patterns. The study of statistics, especially when combined with computer statistics and spreadsheet programs, is a great help in organizing this kind of data, plotting line graphs, and looking for typical underlying relationships. For most economics and social science majors, a statistics course will be required at some point.

Try It

Visit this page in your course online to view this presentation.

One common line graph is called a **time series**, in which the horizontal axis shows time and the vertical axis displays another variable. Thus, a time-series graph shows how a variable changes over time. Figure 3 shows the unemployment rate in the United States since 1975, where unemployment is defined as the percentage of adults who want jobs and are looking for a job, but cannot find one. The points for the unemployment rate in each year are plotted on the graph, and a line then connects the points, showing how the unemployment rate has moved up and down since 1975. With a graph like this, it is easy to spot the times of high unemployment and of low unemployment.

![Figure 3. U.S. Unemployment Rate, 1975–2014.](image)

Try It

### Pie Graphs
A pie graph (sometimes called a pie chart) is used to show how an overall total is divided into parts. A circle represents a group as a whole. The slices of this circular “pie” show the relative sizes of subgroups.

Figure 4 shows how the U.S. population was divided among children, working-age adults, and the elderly in 1970, 2000, and what is projected for 2030. The information is first conveyed with numbers in the table, below, and then in three pie charts.


<table>
<thead>
<tr>
<th>Year</th>
<th>Total Population</th>
<th>19 and Under</th>
<th>20–64 years</th>
<th>Over 65</th>
</tr>
</thead>
<tbody>
<tr>
<td>1970</td>
<td>205.0 million</td>
<td>77.2 (37.6%)</td>
<td>107.7 (52.5%)</td>
<td>20.1 (9.8%)</td>
</tr>
<tr>
<td>2000</td>
<td>275.4 million</td>
<td>78.4 (28.5%)</td>
<td>162.2 (58.9%)</td>
<td>34.8 (12.6%)</td>
</tr>
<tr>
<td>2030</td>
<td>351.1 million</td>
<td>92.6 (26.4%)</td>
<td>188.2 (53.6%)</td>
<td>70.3 (20.0%)</td>
</tr>
</tbody>
</table>

In a pie graph, each slice of the pie represents a share of the total, or a percentage. For example, 50% would be half of the pie and 20% would be one-fifth of the pie. The three pie graphs in Figure 4 show that the share of the U.S. population 65 and over is growing. The pie graphs allow you to get a feel for the relative size of the different age groups from 1970 to 2000 to 2030, without requiring you to slog through the specific numbers and percentages in the table. Some common examples of how pie graphs are used include dividing the population into groups by age, income level, ethnicity, religion, occupation; dividing different firms into categories by size, industry, number of employees; and dividing up government spending or taxes into its main categories.

Bar Graphs

A bar graph uses the height of different bars to compare quantities. The table, below, lists the 12 most populous countries in the world. Figure 5 provides this same data in a bar graph. The height of the bars corresponds to the population of each country. Although you may know that China and India are the most populous countries in the world, seeing how the bars on the graph tower over the other countries helps illustrate the magnitude of the difference between the sizes of national populations.
<table>
<thead>
<tr>
<th>Country</th>
<th>Population</th>
</tr>
</thead>
<tbody>
<tr>
<td>China</td>
<td>1,369</td>
</tr>
<tr>
<td>India</td>
<td>1,270</td>
</tr>
<tr>
<td>United States</td>
<td>321</td>
</tr>
<tr>
<td>Indonesia</td>
<td>255</td>
</tr>
<tr>
<td>Brazil</td>
<td>204</td>
</tr>
<tr>
<td>Pakistan</td>
<td>190</td>
</tr>
<tr>
<td>Nigeria</td>
<td>184</td>
</tr>
<tr>
<td>Bangladesh</td>
<td>158</td>
</tr>
<tr>
<td>Russia</td>
<td>146</td>
</tr>
<tr>
<td>Japan</td>
<td>127</td>
</tr>
<tr>
<td>Mexico</td>
<td>121</td>
</tr>
<tr>
<td>Philippines</td>
<td>101</td>
</tr>
</tbody>
</table>

*Figure 5. The graph shows the 12 countries of the world with the largest populations. The height of the bars in the bar graph shows the size of the population for each country.*

Bar graphs can be subdivided in a way that reveals information similar to that we can get from pie charts. Figure 6 offers three bar graphs based on the information from Figure 4 about the U.S. age distribution in 1970, 2000, and 2030. Figure 6 (a) shows three bars for each year, representing the total number of persons in each age bracket for each year. Figure 6 (b) shows just one bar for each year, but the different age groups are now shaded inside the bar. In Figure 6 (c), still based on the same data, the vertical axis measures percentages rather than the number of persons. In this case, all three bar graphs are the same height, representing 100 percent of the population, with each bar divided according to the percentage of population in each age group. It is sometimes easier for a reader to run his or her eyes across several bar graphs, comparing the shaded areas, rather than trying to compare several pie graphs.
Figure 5 and Figure 6 show how the bars can represent countries or years, and how the vertical axis can represent a numerical or a percentage value. Bar graphs can also compare size, quantity, rates, distances, and other quantitative categories.

Comparing Line Graphs, Pie Charts, and Bar Graphs

Now that you are familiar with pie graphs, bar graphs, and line graphs, how do you know which graph to use for your data? Pie graphs are often better than line graphs at showing how an overall group is divided. However, if a pie graph has too many slices, it can become difficult to interpret.

Bar graphs are especially useful when comparing quantities. For example, if you are studying the populations of different countries, as in Figure 5, bar graphs can show the relationships between the population sizes of multiple countries. Not only can it show these relationships, but it can also show breakdowns of different groups within the population.

A line graph is often the most effective format for illustrating a relationship between two variables that are both changing. For example, time-series graphs can show patterns as time changes, like the unemployment rate over time. Line graphs are widely used in economics to present continuous data about prices, wages, quantities bought and sold, the size of the economy.

Glossary

**bar graph**: shows data as vertical bars; the height of different bars are used to compare quantities
PUTTING IT TOGETHER: ECONOMIC THINKING

The Cost of Waiting in Line

Given your new understanding of economic concepts, let's revisit the example at the beginning of this module: the experience of taking a flight and feeling like others have special privileges that you don't have.

In our example, you waited in line at the security checkpoint for much longer than those who went through the express line. Let's assume that you waited in line for one hour and 10 minutes, while those with express access were able to get through security in just 10 minutes. What is the cost of one hour of your time? What is the cost of a ticket that gets you into the express lane? Did you make the right choice?

Let's return to the concept of opportunity cost. Remember, opportunity cost indicates what must be given up to obtain something that's desired.

You chose to wait in line rather than buying a ticket that would allow you to use the express lane. It may not have seemed like a choice, but you did choose to buy a less expensive ticket—instead of paying more for one would have gotten you into the express line. How much more money would you have to pay for that ticket?

On a flight from Los Angeles to Baltimore, the business select fare is $605. This is the fare that permits access to the express security lane. The lowest fare is $276. The difference is:

\[ \$605 - \$276 = \$329 \]

In other words, you chose to wait in line for one hour in order to save the $329 that you would have had to spend for a business select ticket. When we think of this in terms of opportunity cost, you now have some way to measure your decision: you have a firm number that can be compared against the cost of an alternative.

The Full Cost of Your Time
Let’s consider monetary costs alone. If you had worked at a job for one more hour—instead of waiting in the security line—you could have earned an additional $20. That choice would have meant earning more money, but you would have had to spend far more for a business select ticket than you made in an hour. Since the $329 cost of that ticket is so much greater than the $20 you might have earned, the decision to wait in line for an hour (as opposed to working one hour more) makes good sense when comparing the monetary cost.

But remember, **opportunity cost is the value of the next best alternative**, and there are likely ways that you spend your time that you value more than money. Perhaps getting through security more quickly will enable you to sit down in a quiet café, enjoy a cup of tea, and avoid the stress you feel when you are rushed. There is a real value that you have given up.

Or, what if you are struggling to stay caught up on your schoolwork and an hour of study makes the difference between getting a good grade on an important test or not? If it’s a test that has a big impact on your grade or academic record, then an hour might be incredibly valuable.

In other words, for any individual, the opportunity cost may simply be the lost money from work, or it may be peace of mind, or it may be an hour of study time—or something altogether different. Whichever one of the alternatives has the greatest value to you is your opportunity cost for one hour in line.

**Does It Ever Make Sense to Buy Time?**

Obviously an hour of time has a different opportunity cost for every individual. Let’s take a minute to look at a graph of hourly wages and see if it helps us think about the opportunity cost.

On any flight, passengers with a higher hourly wage are more likely to purchase a select ticket that permits them to bypass security lines and board early. For some groups of passengers, the time is more valuable than the money. Still, you will note that some passengers are willing to spend more than their hourly wage to avoid waiting in line.

The definition of opportunity cost is quite specific: it’s the value of the next best alternative. However, every individual values time, money, convenience, peace of mind, and other factors differently.
Why learn to think like an economist?

As you now know, the study of economics is about choices that are made by individuals and entities, given the fact that we can never have enough. You might not argue that you don’t have enough time or money, for instance, but why might you want to think about that in economic terms? Let’s look at one situation in which the choices you make today—with limited time and money—have an impact on the choices available to you in the future.

It’s generally true that the higher the educational degree a person has, the higher the salary he or she will earn. So why aren’t more people pursuing higher degrees? The short answer: choices and trade-offs.

In 2012, the average annual salary for a full-time U.S. worker over age twenty-five with a master’s degree was $67,600. Compare that to annual earnings for a full-time worker over twenty-five with no higher than a bachelor’s degree: $55,432 a year. What about those with no higher than a high school diploma? They earn just $33,904 in a year. In other words, says the U.S. Bureau of Labor Statistics (BLS), earning a bachelor’s degree boosted salaries 63 percent above what you would have earned if you had stopped your education after high school. A master’s degree yields a salary almost double that of a high school diploma.
What are your educational goals? Do you plan to complete a bachelor’s degree? A master’s degree? Given the salary data, shouldn’t everyone pursue a master’s degree? When you made your own educational plans and goals, perhaps you were motivated by the potential for financial returns later on—I.e., the expectation that a higher degree would lead to a higher-paid job or career. But what other factors did you consider? Perhaps you also thought about the time and cost of education and the other things you like to do when you aren’t studying. Other people, it turns out, also think about these things when deciding whether or not to pursue college.

Considering salary data alone, you might expect a lot of people to choose to attend college and at least earn a bachelor’s degree. In fact, in 2012, the BLS reported that while nearly 88 percent of the U.S. population had a high school diploma, only 31 percent had a bachelor’s degree, and only 8 percent had earned a master’s degree.

For the majority of Americans, the time, money, and effort required to earn a degree is too great, in spite of the resulting salary benefits. In recognition of these barriers, state and federal governments have created programs such as the Pell Grant program to help students pay the financial costs of going to college. However, these programs don’t cover the opportunity costs that are often the most pressing concern for students. For example, the opportunity cost of lost income that could be used to support a student’s family might be a significant factor.

So, now that you’re in college, how can you make the best decision about which level of education to pursue? Perhaps more important, how can you be realistic about your scarce resources and develop a plan that provides the greatest benefit to you?

In this module we will look more closely at the idea of choices and trade-offs, revisit the concept of opportunity cost, and learn how to calculate it. This will help you assign dollar amounts to your choices and understand why your decision to pursue a college degree—in spite of the opportunity costs—is one of the most important decisions you can make toward improving your financial future.

INTRODUCTION TO THE COST OF CHOICES

What you’ll learn to do: explain the cost of choices and trade-offs
In the previous module we introduced the concepts of scarcity, monetary cost, and opportunity cost. This section focuses on the calculation of opportunity costs, or the cost of choosing one thing over another. For example, what alternatives do you give up by choosing to take a new job, or to go out with friends on the weekend? This section will help to understand the consequences of these trade-offs in more concrete terms.

BUDGET CONSTRAINTS AND CHOICES

Learning Objectives

- Explain how budget constraints impact choices

Budget Constraint Framework

For most of us, the idea of scarcity and trade-offs is something we experience in a very real way when it comes to our own budget constraints. Most of us have a limited amount of money to spend on the things we need and want. Another kind of budget constraint is time. For instance, as a student, you only have twenty-four hours in the day to study, eat, sleep, and check Facebook. An hour spent studying economics is an hour that can't be used for sleep or play (or something else). As a result, you have to make choices, and every choice involves trade-offs.

In economics, a **budget constraint** refers to all possible combinations of goods that someone can afford, given the prices of goods and the income (or time) we have to spend.
Take the following example of someone who must choose between two different goods: Charlie has $10 in spending money each week that he can allocate between bus tickets for getting to work and the burgers he eats for lunch. Burgers cost $2 each, and bus tickets are 50 cents each. Figure 2, below, shows Charlie’s budget constraint ($10) and all the possible combinations of burgers and bus tickets he can afford if he spends all his money.

![Figure 2. Charlie’s budget constraint.](image)

The vertical axis in the figure shows burger purchases, and the horizontal axis shows bus ticket purchases. If Charlie spends all his money on burgers, he can afford five per week. ($10 per week/$2 per burger = 5 burgers per week.) But if he does this, he won’t be able to afford any bus tickets. This choice (zero bus tickets and 5 burgers) is shown by point A in the figure. Alternatively, if Charlie spends all his money on bus tickets, he can afford 20 per week. ($10 per week/$0.50 per bus ticket = 20 bus tickets per week.) Then, however, he will not be able to afford any burgers. This alternative choice (20 bus tickets and zero burgers) is shown by point F.

If Charlie is like most people, he will choose some combination that includes both bus tickets and burgers—that is, he will choose one of the points along the budget-constraint line that connects points A and F. Each point inside or on the budget constraint shows a combination of burgers and bus tickets that Charlie can afford. (A point inside the curve is definitely an option—it just means that Charlie isn’t spending all his money.) Keep in mind that the curve represents the maximum number of burgers and bus tickets he can buy. Any point outside the constraint is not affordable, because it would cost more money than Charlie has in his budget. The slope of the budget constraint is determined by the relative price of burgers and bus tickets.

The budget constraint clearly shows the trade-off Charlie faces in choosing between burgers and bus tickets. Suppose he is currently at point D, where he can afford 12 bus tickets and 2 burgers. What would it cost Charlie for one more burger? It would be natural to answer $2, but that’s not the way economists think. Instead, they ask: how many bus tickets would Charlie have to give up to get one more burger, while staying within his budget? The answer is four bus tickets. That is the true cost to Charlie of one more burger.

A budget-constraint diagram like the one above, with just two goods—burgers and bus tickets—is simple and not very realistic. After all, in an economy like ours (and Charlie’s), people choose from thousands of goods. However, no matter how many goods a consumer has to choose from, every choice has an opportunity cost, i.e. the value of the other goods that aren’t chosen. This is the point that carries over to the real world.

**Try It**

Visit this page in your course online to view this presentation.

**Sunk Costs**

In the budget constraint framework, all decisions involve what will happen next: What quantities of goods will you consume? How many hours will you work? How much will you save? Choices made or costs in the past are not taken into account. The budget constraint framework assumes that sunk costs—costs incurred in the past that can’t be recovered—should not affect the current decision.
Suppose you pay $8 to see a movie, but after watching the first thirty minutes, you decide that it's awful. Should you stick it out and watch the rest because you paid for the ticket, or should you leave? The money you spent on the ticket is a sunk cost, and unless the theater manager is feeling generous, you won't get a refund. But staying for the rest of the movie means paying an opportunity cost in time. Your choice is whether to spend the next ninety minutes suffering through a rotten movie or do something—anything—else. The lesson of sunk costs is to forget about the money and time that is irretrievably gone and to focus, instead, on the costs and benefits of current and future options. A sunk cost is water under the bridge, so to speak.

For people and organizations alike, dealing with sunk costs can be frustrating and difficult. For one thing, it often means admitting an earlier error of judgment. Many companies find it hard to give up on a new product that's doing poorly because they've invested so much time and money in the product development and launch. But the lesson of sunk costs is to ignore them and make decisions based on what will happen in the future.

Glossary

**budget constraint**: all possible consumption combinations of goods that someone can afford, given the prices of goods, when all income is spent; the boundary of the opportunity set

**opportunity cost**: measures cost by what is given up in exchange; opportunity cost measures the value of the forgone alternative

**sunk costs**: costs that are made in the past and cannot be recovered

Learning Objectives

- Calculate the opportunity costs of an action
It makes intuitive sense that Charlie can buy only a limited number of bus tickets and burgers with a limited budget. Also, the more burgers he buys, the fewer bus tickets he can buy. With a simple example like this, it isn’t too hard to determine what he can do with his very small budget, but when budgets and constraints are more complex, equations can be used to demonstrate budget constraints and opportunity cost.

Very simply, when Charlie is spending his full budget on burgers and tickets, his budget is equal to the total amount that he spends on burgers plus the total amount that he spends on bus tickets. For example, if Charlie buys four bus tickets and four burgers with his $10 budget (point B on the graph below), the equation would be

\[ \$10 = (\$2 \times 4) + (\$0.50 \times 4) \]

You can see this on the graph of Charlie’s budget constraint, Figure 1, below.

*Figure 1. Charlie’s Budget Constraint.*

If we want to answer the question, “how many burgers and bus tickets can Charlie buy?” then we need to use the budget constraint equation.

**Work It Out**

**Step 1.** The equation for any budget constraint is the following:

\[
\text{Budget} = P_1 \times Q_1 + P_2 \times Q_2 + \cdots + P_n \times Q_n
\]

where \( P \) and \( Q \) are the price and respective quantity of any number, \( n \), of items purchased and Budget is the amount of income one has to spend.

**Step 2.** Apply the budget constraint equation to the scenario.

In Charlie’s case, this works out to be

\[
\text{Budget} = P_1 \times Q_1 + P_2 \times Q_2
\]

Budget = $10

\[
P_1 = \$2 \text{ (the price of a burger)}
\]

\[
Q_1 = \text{quantity of burgers (variable)}
\]

\[
P_2 = \$0.50 \text{ (the price of a bus ticket)}
\]
$Q_2 = \text{quantity of tickets (variable)}$

For Charlie, this is:

$10 = 2Q_1 + 0.50Q_2$

**Step 3.** Simplify the equation.

At this point we need to decide whether to solve for $Q_1$ or $Q_2$.

Remember, $Q_1 = \text{quantity of burgers}$. So, in this equation $Q_1$ represents the number of burgers Charlie can buy depending on how many bus tickets he wants to purchase in a given week. $Q_2 = \text{quantity of tickets}$. So, $Q_2$ represents the number of bus tickets Charlie can buy depending on how many burgers he wants to purchase in a given week.

We are going to solve for $Q_1$.

\[
10 = 2Q_1 + 0.50Q_2 \\
10 - 2Q_1 = 0.50Q_2 \\
-2Q_1 = -10 + 0.50Q_2 \\
(2)(-2Q_1) = (2) - 10 + (2)(0.50Q_2) \quad \text{Clear decimal by multiplying everything by 2} \\
-4Q_1 = -20 + Q_2 \\
Q_1 = 5 - \frac{1}{4}Q_2 \quad \text{Divide both sides by $-4$}
\]

**Step 4.** Use the equation.

Now we have an equation that helps us calculate the number of burgers Charlie can buy depending on how many bus tickets he wants to purchase in a given week.

For example, say he wants 8 bus tickets in a given week. $Q_2$ represents the number of bus tickets Charlie buys, so we plug in 8 for $Q_2$, which gives us

\[
Q_1 = 5 - \frac{1}{4}(8) \\
Q_1 = 5 - 2 \\
Q_1 = 3
\]

This means Charlie can buy 3 burgers that week (point C on the graph, above).

Let's try one more. Say Charlie has a week when he walks everywhere he goes so that he can splurge on burgers. He buys 0 bus tickets that week. $Q_2$ represents the number of bus tickets Charlie buys, so we plug in 0 for $Q_2$, giving us

\[
Q_1 = 5 - \frac{1}{4}(0) \\
Q_1 = 5
\]

So, if Charlie doesn't ride the bus, he can buy 5 burgers that week (point A on the graph).

If you plug other numbers of bus tickets into the equation, you get the results shown in Table 1, below, which are the points on Charlie's budget constraint.

<table>
<thead>
<tr>
<th>Point</th>
<th>Quantity of Burgers (at $2)</th>
<th>Quantity of Bus Tickets (at 50 cents)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>C</td>
<td>3</td>
<td>8</td>
</tr>
</tbody>
</table>
Table 1.

<table>
<thead>
<tr>
<th>Point</th>
<th>Quantity of Burgers (at $2)</th>
<th>Quantity of Bus Tickets (at 50 cents)</th>
</tr>
</thead>
<tbody>
<tr>
<td>D</td>
<td>2</td>
<td>12</td>
</tr>
<tr>
<td>E</td>
<td>1</td>
<td>16</td>
</tr>
<tr>
<td>F</td>
<td>0</td>
<td>20</td>
</tr>
</tbody>
</table>

**Step 4.** Graph the results.

If we plot each point on a graph, we can see a line that shows us the number of burgers Charlie can buy depending on how many bus tickets he wants to purchase in a given week.

![Budget Constraint Graph](image)

*Figure 2. Charlie's Budget Constraint.*

We can make two important observations about this graph. First, the slope of the line is negative (the line slopes downward from left to right). Remember in the last module when we discussed graphing, we noted that when when X and Y have a negative, or inverse, relationship, X and Y move in opposite directions—that is, as one rises, the other falls. This means that the only way to get more of one good is to give up some of the other.

Second, the slope is defined as the change in the number of burgers (shown on the vertical axis) Charlie can buy for every incremental change in the number of tickets (shown on the horizontal axis) he buys. If he buys one less burger, he can buy four more bus tickets. The slope of a budget constraint always shows the opportunity cost of the good that is on the horizontal axis. If Charlie has to give up lots of burgers to buy just one bus ticket, then the slope will be steeper, because the opportunity cost is greater.

This is easy to see while looking at the graph, but opportunity cost can also be calculated simply by dividing the cost of what is given up by what is gained. For example, the opportunity cost of the burger is the cost of the burger divided by the cost of the bus ticket, or

\[
\frac{2.00}{0.50} = 4
\]

The opportunity cost of a bus ticket is:

\[
\frac{0.50}{2.00} = 0.25
\]

Let's look at this in action and see it on a graph. What if we change the price of the burger to $1? We will keep the price of bus tickets at 50 cents.
LEARN BY DOING: BUDGET CONSTRAINTS AND OPPORTUNITY COST

Try It

Rather than have you read more about budget constraints and opportunity costs, we’d prefer to have you practice what you’ve done so far and see for yourself if you understand it.

These questions allow you to get as much practice as you need, as you can click the link at the top of the first question (“Try another version of these questions”) to get a new version of the questions. Practice until you feel comfortable doing these questions.

Note that you’ll use the information provided in the first question for all of the questions on this page.

Visit this page in your course online to practice before taking the quiz.

INTRODUCTION TO THE PRODUCTION POSSIBILITIES FRONTIER

What you’ll learn to do: illustrate society’s trade-offs by using a production possibilities frontier (or curve)
In the previous sections of this module, we explored how individuals make choices about how to spend their budgets. In this section, we expand that idea to look at how societies make choices about what goods and services to produce.

During the 1960s, President Lyndon Johnson attempted two major, costly initiatives: launching the “War on Poverty,” and expanding the Vietnam War. The results were not as nice as he expected because he was unable to completely tackle both issues, and he ignored what economists call the production possibilities frontier (also called the production possibilities curve). In brief, societies have limited resources so they face trade-offs, just as individuals do.

The production possibilities frontier (or PPF, for short) is a model of the economy as a whole, which shows all possible combinations of goods products or services that a society could produce, given the resources it has available.

Learning Objectives

- Explain the production possibilities frontier

Just as individuals cannot have everything they want and must instead make choices, society as a whole cannot have everything it might want, either. Economists use a model called the production possibilities frontier (PPF) to explain the constraints society faces in deciding what to produce.
As you read this section, you will see parallels between individual choice and societal choice. There are more similarities than differences, so for now focus on the similarities.

While individuals face budget and time constraints, societies face the constraint of limited resources (e.g., labor, land, capital, raw materials, etc.). Because at any given moment, society has limited resources, it follows that there’s a limit to the quantities of goods and services it can produce. In other words, the products are limited because the resources are limited.

Suppose a society desires two products: health care and education. This situation is illustrated by the production possibilities frontier in Figure 1.

Health care is shown on the vertical (or y) axis, and education is shown on the horizontal (or x) axis. Where does the PPF come from? It comes from the production processes for producing the two goods, and the limited amounts of resources available to use for that purpose. For example, suppose one teacher can teach 25 students in school. If society has a total of 10 teachers, education can be provided to a maximum of 250 students. We would say one teacher could “produce” 25 students worth of education using the education processes available.

Suppose a society allocated all of its resources to producing health care. That is certainly one possible way of allocating a society’s resources, but it would mean there would be no resources left for education. This choice is shown in Figure 1 at point A. Similarly, the society could allocate all of its resources to producing education, and none to producing healthcare, as shown at point F. Alternatively, the society could choose to produce any combination of health care and education shown on the production possibilities frontier. In effect, the production possibilities frontier plays the same role for society as the budget constraint plays for an individual consumer. Society can choose any combination of the two goods on or inside the PPF. However, it does not have enough resources to produce outside the PPF.

Most importantly, the production possibilities frontier clearly shows the tradeoff between healthcare and education. Suppose society has chosen to operate at point B, and it’s considering producing more education. Because the PPF is downward sloping from left to right, the only way society can obtain more education is by giving up some health care. That’s the trade-off this society faces. Suppose it considers moving from point B to point C. What would be the opportunity cost for the additional education? The opportunity cost would be the health care that society has to give up.

Do you remember Charlie choosing combinations of burgers and bus tickets within his budget constraint? In effect, the production possibilities frontier plays the same role for society as the budget constraint plays for Charlie. Society can choose any combination of the two goods on or inside the PPF, but it doesn’t have enough resources to produce outside the PPF. Just as with Charlie’s budget constraint, the opportunity cost is shown by the slope of the production possibilities frontier.

Try It

Visit this page in your course online to view this presentation.
Differences between a Budget Constraint and a PPF

We're now ready to address the differences between society's PPF and an individual's budget constraint.

A budget constraint shows the different combinations of goods and services a consumer can purchase with their fixed budget. A production possibilities frontier shows the possible combinations of goods and services that a society can produce with its limited resources. The first difference between a budget constraint and a production possibilities frontier is that the PPF, because it's looking at societal choice, is going to have much larger numbers on the axes than those on an individual's budget constraint.

The most important difference between the two graphs, though, is that a budget constraint is a straight line, while a production possibilities curve is typically bowed outwards, i.e. concave towards the origin. The reason for this difference is pretty simple: the slope of a budget line is defined as the ratio of the prices of the two goods or services. No matter how many of each good or service a consumer buys, the prices stay the same. By contrast, the slope of a PPF is the cost to society of producing one good or service relative to the other good or service. When society reallocates resources from one product to another, the relative costs change, which means the slope of the PPF does also. Let’s dig into this.

To understand why the PPF is curved, start by considering point A at the top left-hand side of the PPF. Considering the situation in Figure 1 (shown again below), suppose we have only two types of resources: doctors and teachers. At point A, all available resources (i.e. all the doctors and all the teachers) are devoted to providing health care and none is left for education. Say the doctors are practicing medicine and the teachers are helping out as best they can. This situation would be extreme and even ridiculous. For example, children are seeing a doctor every day, whether they’re sick or not, but not attending school. People are having cosmetic surgery on every part of their bodies, but no high school or college education exists!

Now imagine that some of these resources are diverted from health care to education, so that the economy is at point B instead of point A. What type of resources are going to move to producing education? The doctors are good at medicine, but they're not particularly good at teaching, so it doesn't make sense for them to switch. The teachers, though, are good at education, and not very good at healthcare. After all, that's not what they were trained for. So it makes sense for teachers to be reallocated from healthcare to education. And when they move, the society doesn't lose much healthcare, because the teachers weren't very good at that. But the amount of education gained is great, because that's what teachers are trained for. What this means is that from point A to B, the decrease in healthcare is small, while the gain in education is large. Graphically, the rise is small and the run is large so the slope (which is the ratio of rise over run) is flat. In other words, the opportunity cost of education in terms of healthcare is low.
If we started at the other end of the PPF at point F and moved to point D, we would be moving doctors from teaching to healthcare with the result that the gain in healthcare would be large while the loss in education would be small (the same logic we used above). In short, the slope of the PPF from point F to D would be steep, and the opportunity cost of education in terms of healthcare would be high.

More generally, as society produces more and more of some good or service, the cost of production grows larger and larger relative to the cost of producing other goods or services. Thus, the slope of a PPF starts flat and becomes increasingly steeper. In the real world, of course, we have more than two goods and services, and we have more resources than just labor, but the general rule still holds.

There's another way to think about this. For consumers, there is only one scarce resource: budget dollars. As we choose more of one good and less of another, we are simply spending dollars on different items, but every dollar is worth the same in purchasing any item. For society, there are many scarce resources. In our simple example above, there were two different resources: doctors and teachers, and each resource is better at one job than at the other. In other words, each resource is not worth the same at producing different products. The general rule is when one is allocating only a single scarce resource, the trade-off (e.g. budget line) will be constant, but when there is more than one scarce resources, the trade-off will be increasingly costly (e.g. the PPF).

The Law of Diminishing Returns and the Curved Shape of the PPF

The lesson is not that society is likely to make an extreme choice like devoting no resources to education at point A or no resources to health at point F. Instead, the lesson is that the gains from committing additional marginal resources to education depend on how much is already being spent. If, on the one hand, very few resources are currently committed to education, then an increase in resources used can bring relatively large gains. On the other hand, if a large number of resources are already committed to education, then committing additional resources will bring relatively smaller gains.

This pattern is so common that it has been given a name: the law of diminishing returns. This law asserts that as additional increments of resources are devoted to a certain purpose, the marginal benefit from those additional increments will decline. For example, after not spending much at all on crime reduction, when a government spends a certain amount more, the gains in crime reduction could be relatively large. But additional increases after that typically cause relatively smaller reductions in crime, and paying for enough police and security to reduce crime to zero would be tremendously expensive.
The curve of the production possibilities frontier shows that as additional resources are added to education, moving from left to right along the horizontal axis, the initial gains are fairly large, but those gains gradually diminish. Similarly, as additional resources are added to health care, moving from bottom to top on the vertical axis, the initial gains are fairly large but again gradually diminish. In this way, the law of diminishing returns produces the outward-bending shape of the production possibilities frontier.

Try It
Visit this page in your course online to view this presentation.

Glossary

**law of diminishing returns**: as additional increments of resources are devoted to a certain purpose, the marginal benefit from those additional increments will decline

**production possibilities frontier (or curve)**: a diagram that shows the productively efficient combinations of two products that an economy can produce given the resources it has available
PRODUCTIVE EFFICIENCY AND ALLOCATIVE EFFICIENCY

Learning Objectives
- Use the production possibilities frontier to identify productive and allocative efficiency

Efficiency

In a market-oriented economy with a democratic government, the choice of what combination of goods and services to produce, and thus where to operate along the production possibilities curve, will involve a mixture of decisions by individuals, firms, and government, expressing supplies and demands. However, economics can point out that some choices are unambiguously better than others. This observation is based on the idea of efficiency. In everyday parlance, *efficiency* refers to lack of waste. An inefficient washing machine operates at high cost, while an efficient washing machine operates at lower cost, because it’s not wasting water or energy. An inefficient organization operates with long delays and high costs, while an efficient organization is focused, meets deadlines, and performs within budget.

The production possibilities frontier can illustrate two kinds of efficiency: productive efficiency and allocative efficiency. Figure 2, below, illustrates these ideas using a production possibilities frontier between health care and education.

![Figure 1. Laundry Costs. An efficient washing machine operates at a low cost.](image)

![Figure 2. Productive and Allocative Efficiency. Points along the PPF display productive efficiency while those point R does not.](image)

**Productive efficiency** means that, given the available inputs and technology, it’s impossible to produce more of one good without decreasing the quantity of another good that’s produced. All choices along the PPF in Figure 2, such as...
points A, B, C, D, and F, display productive efficiency. As a firm moves from any one of these choices to any other, either health care increases and education decreases or vice versa. This makes sense if you remember the definition of the PPF as showing the maximum amounts of goods a society can produce, given the resources it has. Thus, producing efficiently leads to maximum production, which is what the PPF shows.

However, any choice inside the production possibilities frontier is productively inefficient and wasteful because it's possible to produce more of one good, the other good, or some combination of both goods. Wasting scarce resources means the society is not producing as well or as much as it could, so it is not operating on the PPF.

For example, point R is productively inefficient because it is possible at choice C to have more of both goods: education on the horizontal axis is higher at point C than point R (E₂ is greater than E₁), and health care on the vertical axis is also higher at point C than point R (H₂ is greater than H₁).

Any time a society is producing a combination of goods that falls along the PPF, it is achieving productive efficiency. When the combination of goods produced falls inside the PPF, then the society is productively inefficient.

Allocative efficiency means that the particular mix of goods a society produces represents the combination that society most desires. For example, often a society with a younger population has a preference for production of education, over production of health care. If the society is producing the quantity or level of education that the society demands, then the society is achieving allocative efficiency. Determining “what a society desires” can be a controversial question and is often discussed in political science, sociology, and philosophy classes, as well as in economics.

At the most basic level, allocative efficiency means that producers supply the quantity of each product that consumers demand. Only one of the productively efficient choices will be the allocative efficient choice for society as a whole. For example, in order to achieve allocative efficiency, a society with a young population will invest more in education. As the population ages, the society will shift resources toward health care because the older population requires more health care than education.

In the graph (Figure 2) above, a society with a younger population might achieve allocative efficiency at point D, while a society with an older population that required more health care might achieve allocative efficiency at point B.

We will return to this idea of allocative efficiency later when we learn more about applications of supply and demand.

Try It

Visit this page in your course online to view this presentation.

Why Society Must Choose

Every economy faces two situations in which it may be able to expand the consumption of all goods. In the first case, a society may discover that it has been using its resources inefficiently, in which case by improving efficiency and producing on the production possibilities frontier, it can have more of all goods (or at least more of some and less of none). In the second case, as resources grow over a period of years (e.g., more labor and more capital), the economy grows. As it does, the production possibilities frontier for a society will tend to shift outward, and society will be able to afford more of all goods.

However, improvements in productive efficiency take time to discover and implement, and economic growth happens only gradually. So, a society must choose between trade-offs in the present—as opposed to years down the road. For government, this process often involves trying to identify where additional spending could do the most good and where reductions in spending would do the least harm. At the individual and firm level, the market economy coordinates a process in which firms seek to produce goods and services in the quantity, quality, and price that people want. But for both the government and the market economy, in the short term, increases in production of one good typically mean offsetting decreases somewhere else in the economy.
The PPF and Comparative Advantage

While every society must choose how much of each good it should produce, it doesn’t need to produce every single good it consumes. Often, how much of a good a country decides to produce depends on how expensive it is to produce it versus buying it from a different country. As we saw earlier, the curve of a country’s PPF gives us information about the trade-off between devoting resources to producing one good versus another. In particular, its slope gives the opportunity cost of producing one more unit of the good in the x-axis in terms of the other good (in the y-axis). Countries tend to have different opportunity costs of producing a specific good, either because of different climates, geography, technology, or skills.

Suppose two countries, the U.S. and Brazil, need to decide how much they will produce of two crops: sugar cane and wheat. Due to its climate, Brazil can produce a lot of sugar cane per acre but not much wheat. Conversely, the U.S. can produce a lot of wheat per acre, but not much sugar cane. Clearly, Brazil has a lower opportunity cost of producing sugar cane (in terms of wheat) than the U.S. The reverse is also true; the U.S. has a lower opportunity cost of producing wheat than Brazil. This can be illustrated by the PPF of each country, shown in Figure 4, below.

![Figure 4. Brazil and U.S. production possibility frontiers.](image)

When a country can produce a good at a lower opportunity cost than another country, we say that this country has a **comparative advantage** in that good. In our example, Brazil has a comparative advantage in sugar cane, and the U.S. has a comparative advantage in wheat. One can easily see this with a simple observation of the extreme production points in the PPFs. If Brazil devoted all of its resources to producing wheat, it would be producing at point A. If, however, it devoted all of its resources to producing sugar cane instead, it would be producing a much larger amount, at point B. By moving from point A to point B, Brazil would give up a relatively small quantity in wheat production to obtain a large production in sugar cane. The opposite is true for the U.S. If the U.S. moved from point A to B and produced only sugar cane, this would result in a large opportunity cost in terms of foregone wheat production.
The slope of the PPF gives the opportunity cost of producing an additional unit of wheat. While the slope is not constant throughout the PPFs, it is quite apparent that the PPF in Brazil is much steeper than in the U.S., and therefore the opportunity cost of wheat is generally higher in Brazil. In the module on international trade you will learn that countries’ differences in comparative advantage determine which goods they will choose to produce and trade. When countries engage in trade, they specialize in the production of the goods in which they have comparative advantage and trade part of that production for goods in which they don’t have comparative advantage in. With trade, goods are produced where the opportunity cost is lowest, so total production increases, benefiting both trading parties.

Glossary

**allocative efficiency**: when the mix of goods being produced represents the mix that society most desires

**productive efficiency**: given the available inputs and technology, it’s impossible to produce more of one good without decreasing the quantity of another good that’s produced

INTRODUCTION TO THE ECONOMIC WAY OF THINKING

What you’ll learn to do: explain the assumption of economic rationality, define marginal analysis, and differentiate between positive and normative reasoning.
This module is about how economists analyze issues and problems, which is sometimes referred to as the “economic way of thinking.” In the previous sections of the module, we explored two common models used by economists to think about economic issues. Now we segue into introducing some specific features of economic thinking: economic rationality, marginal analysis, and positive vs. normative reasoning.

Economists assume that humans make decisions in predictable ways. They believe that, when making choices, people try to avoid costs and maximize benefits to themselves. This is what economists mean by rational decision-making.

Economists recognize that very few choices in the real world are “all or nothing.” Most of the time, people have the choice to do a little more or a little less of something: Should you eat one more muffin? Should you study economics for another hour? Should you spend a little less money on gas? Economists use the word marginal to mean “additional” or “extra,” and they use the term marginal analysis to describe how people make choices by comparing the benefits and costs of doing a bit more or a bit less.

Economists can make two kinds of arguments. Positive reasoning is scientific reasoning, based on theories and evidence. Policy decisions often employ normative reasoning, which is based on values. For reasons we will see later, it is important to be able to to differentiate between the two.

RATIONALITY AND SELF-INTEREST

Learning Objectives

- Define rationality in an economic context
- Provide examples of rational decision-making
If you say that someone is behaving “rationally,” you probably mean that he or she is acting in a thoughtful, clear-headed way (as opposed to irrationally, which suggests that someone is acting emotionally or illogically). In the context of economics, the term rationality has a very specific meaning. It refers to an assumption that economists make about how people behave—remember that this is the starting point of all economics—in the face of scarcity. There simply aren’t enough resources to satisfy all needs and wants. Charlie has only $10, he’s hungry, and he needs to get to work. What will he do? An economist predicts that Charlie will behave in a predictable, rational manner, balancing costs against benefits to arrive at an action that maximizes his personal happiness or utility. As a result, he will choose a certain number of burgers and a certain number of bus tickets.

To put it differently, if an individual acts in an economically rational way, anything that increases the benefits or decreases the costs of some action is likely to increase the probability that the individual will choose that action. Anything that decreases the benefits or increases the costs will likely reduce the probability that the individual will choose that action.

Economists assume that people will make choices in their own self-interest. They will choose those things that provide the greatest personal benefit, and they’ll avoid or forego those that aren’t as personally valuable and compelling. That’s what we mean by the assumption of rationality.

Do economists really believe that we only think of ourselves and don’t ever try to benefit others? Not at all. The assumption that individuals are purely self-interested doesn’t imply that individuals are greedy and selfish. People clearly derive satisfaction from helping others, so “self-interest” can also include pursuing things that benefit other people. One example is charitable behavior. When National Public Radio holds a fund drive, they often announce a situation where a benefactor has agreed to match the contributions made over a certain time period. If you pledge money during that time period, your action raises twice the amount you contribute. These matching fund situations tend to increase the amount of contributions, because people respond rationally, even when they are giving money to charity.

The assumption of rationality—also called the theory of rational behavior—is primarily a simplification that economists make in order to create a useful model of human decision-making.

**Try It**

Visit this page in your course online to view this presentation.

If you consider your own personal choices, you will probably find that they are quite complex. You are balancing what you want right now with options you want to have in the future. You probably value the people around you—friends, family, neighbors—and you may consider the impact that your choices have on them.

Setting aside the messy realm of personal choices for the time being, let’s take a look at how decisions are made by consumers, by students, and by businesses in a world of economic rationality.

**Rationality and Consumers**

When a consumer is thinking about buying a product, what does he or she want? The theory of rational behavior would say that the consumer wants to maximize benefit and minimize cost.
Figure 2. Is it worth the cost to see a movie in the theater, or would you prefer to wait until you can buy the DVD or watch it on Netflix?

Let's look at a simple example. When a new movie is released, will you see it in the theater, or will you wait for it to be released on Netflix or on TV? If we consider only the monetary costs of your choice, a movie ticket might cost $10 and you will only be able to see that movie one time. If you wait, you can probably watch it as part of your monthly Netflix or cable subscription without spending any more than you would spend without watching the movie.

Why would you pay $10 to watch the movie in the theater? You might want to see it right away, when it is only showing in the theater. You might want the theater experience, with the big screen and high-quality image and sound. You will make a decision that is economically rational, based on the following consideration: “Is the benefit and enjoyment that I get from seeing the movie in a theater worth the $10 cost?”

As a consumer, you are making an economically rational decision about the costs and benefits.

Since we will build upon this later in the course, it's important to understand that this assumption creates a link between the cost of a product and the degree to which a consumer will want to buy it. As the cost of the product increases, it becomes less likely that the consumer will decide that the benefits of the purchase outweigh the costs.

Rationality and Students

How do students decide on a major? A number of things may factor into the decision, such as what type of career a student is interested in, the reputation of specific departments at the university a student is attending, and the student's preferences for specific fields of study. Let's take an example.

You go to college with the idea that you want to major in Business Management. During your first year, you discover that Business Analytics majors earn significantly higher salaries. This discovery increases the benefits in your mind of the Analytics major, and you decide to choose that major. You've just made an economically rational decision.

Rationality and Businesses

Businesses also have predictable behavior, but rather than seeking to maximize happiness or pleasure, they seek to maximize profits. When economists assume that businesses have a goal of maximizing profits, they can make predictions about how companies will react to changing business conditions.

For example, if wages in the United States increase, how will U.S. companies react? The rational reaction may be to move those jobs that can be performed elsewhere to countries with lower wages. This prediction is based on an oversimplification, and it might not hold true in every case—individual businesses would obviously need to understand the full cost of moving certain work out of the country before doing so. But the decision would be made according to the impact on profit and would still be rational. If a company stands to earn more profit by moving some jobs overseas, then that's the result that economists would predict.

Rationality suggests that consumers will act to maximize self-interest and businesses will act to maximize profits. Both are taking into account the benefits of a choice, given the costs.
Glossary

**assumption of rationality**: also called the theory of rational behavior, it is the assumption that people will make choices in their own self-interest

---

**MARGINAL ANALYSIS**

Learning Objectives

- Explain the importance of marginal analysis in economics
- Give examples of marginal cost and marginal benefit

A Little More or a Little Less

The budget constraint framework helps to illustrate that most choices in the real world are not about getting all of one thing or all of another—we rarely decide “all burgers” or “all bus tickets.” Options usually fall somewhere on a continuum, and the choice usually involves marginal decision-making and marginal analysis.

Marginal decision-making means considering a little more or a little less than what we already have. We decide by using **marginal analysis**, which means comparing the costs and benefits of a little more or a little less.

It’s natural for people to compare costs and benefits, but often we look at total costs and total benefits, when the best choice requires comparing how costs and benefits change from one option to another. In short, you might think of marginal analysis as “change analysis.” Marginal analysis is used throughout economics. This subtle concept is easier to grasp with examples.

**Marginal Cost**

Generally speaking, **marginal cost** is the difference (or change) in cost of a different choice. From a consumer’s point of view, marginal cost is the additional cost of one more item purchased. From a business’s point of view, marginal cost is the additional cost of one more item produced.

Suppose you typically spend a week at the beach for vacation, but this year you earned an annual bonus from your job. Should you rent a beach house for one week or two? A one-week rental costs $2,000. A two-week rental costs $3,600. Holding everything else constant, which option is better? If you stay for two weeks, the cost is significantly higher: $3,600 versus $2,000. But consider the cost by week. The first week costs $2,000. The difference in cost between one week and two is $3,600 – $2,000, or $1,600. Thus, while the marginal cost of the first week’s rental is $2,000, the marginal cost of the second week’s rental is $1,600. This illustrates the key rule of marginal analysis: Marginal cost = the change in total cost from one option to another.

Consider another example. Imagine that you’re out getting ice cream with your friends or family. You can choose whether to buy one, two, or three scoops of ice cream. One scoop costs $3.00, two scoops cost $5.00, and three
scoops cost $7.00. This information is shown in the following table.

<table>
<thead>
<tr>
<th>Scoops of Ice Cream</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Cost</td>
<td>$3</td>
<td>$5</td>
<td>$7</td>
</tr>
</tbody>
</table>

What is the marginal cost of each scoop of ice cream? The marginal cost of the first scoop of ice cream is $3.00 because you have to pay $3.00 more to get one scoop of ice cream than you do to get zero scoops of ice cream. The marginal cost of the second scoop of ice cream is $2.00 because you only need to pay two more dollars to get two scoops than you need to pay to get one scoop. The marginal cost of the third scoop is also $2.00 because you would need to pay an additional two dollars to get that third scoop.

<table>
<thead>
<tr>
<th>Scoops of Ice Cream</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marginal Cost</td>
<td>$3</td>
<td>$2</td>
<td>$2</td>
</tr>
</tbody>
</table>

Marginal costs sometimes go up and sometimes go down, but to get the clearest view of your options, you should always try to make decisions based on marginal costs, rather than total costs.

Marginal Benefit

Generally speaking, **marginal benefit** is the difference (or change) in what you receive from a different choice. From a consumer’s point of view, marginal benefit is the additional satisfaction of one more item purchased. From a business’ point of view, marginal benefit is the additional revenues received from selling one more item.

Suppose you’re considering membership at the local recreation center. The basic membership gives access to the swimming pool, while the full membership gives access to the swimming pool and the weight room. What is the difference between the two memberships? Since both give access to the pool, the marginal benefit of full membership is access to the weight room.

The amount of benefit a person receives from a particular good or service is subjective; one person may get more satisfaction or happiness from a particular good or service than another. For example, you might enjoy ice cream more than your friend who is allergic to dairy. The amount of benefit you get can also change. For example, you might enjoy the ice cream more on a hot day than on a cold day. This doesn’t make it any less real, however.
Economic Rationality Revisited

How, then, do you decide on a choice? The answer is that you compare, to the best of your ability, the marginal benefits with the marginal costs. An economically rational decision is one in which the marginal benefits of a choice are greater than the marginal costs of the choice.

If we return to the recreation center example above, suppose that the basic membership is $30 per month, while the full membership is $40 per month. An economically rational decision-maker would ask, Is the marginal benefit (access to the weight room) worth the marginal cost (an extra $10 per month)? For some people, the answer will be yes. For others, it will be no. Either way, marginal analysis is an important part of economic rationality and good decision-making.

Try It

Visit this page in your course online to view this presentation.

Try It

These next questions allow you to get as much practice as you need, as you can click the link at the top of the first question (“Try another version of this question”) to get a new set of questions. Practice until you feel comfortable doing the questions and then move on.

Visit this page in your course online to practice before taking the quiz.

Glossary

**marginal analysis**: examination of decisions on the margin, meaning comparing costs of a little more or a little less

**marginal benefit**: the difference (or change) in what you receive from a different choice

**marginal cost**: the difference (or change) in cost of a different choice

Learning Objectives

- Distinguish between positive and normative statements
Economists engage in two distinct, but related activities. They conduct research on economic issues, e.g. to determine cause and effect. For example, why did unemployment increase rapidly in 2008 and 2009? Economists also make policy recommendations. For example, what should the federal government do in response to the increase in unemployment?

The first type of activity is economic science, based on theories and evidence, where researchers attempt to determine how the world (or at least the economy) works. This is called positive reasoning, and the conclusions are called positive statements. A relevant conclusion might be that because the level of employment is based on production in the economy (i.e. GDP), the increase in unemployment was because of the slowdown in GDP over that time period. This slowdown has been called the Great Recession.

The second type of activity is more subjective, and is inevitably based on the researcher’s values. This is called normative reasoning, and the conclusions are called normative statements. A policy recommendation could be that since unemployed workers are not earning income, government should try to stimulate demand in the economy, so unemployed workers could get back to work. A different policy recommendation could be that stimulating demand could involve running a larger federal budget deficit, which future generations would have to pay back through higher taxes, so the government shouldn’t try to stimulate demand. Which of these recommendations is the right one? That depends on your subjective values.

Positive Statements

Positive statements (and positive reasoning more generally) are objective. As such, they can be tested. These fall into two categories. One is a hypothesis, like “unemployment is caused by a decrease in GDP.” This claim can be tested empirically by analyzing the data on unemployment and GDP. The other category is a statement of fact, such as “It’s raining,” or “Microsoft is the largest producer of computer operating systems in the world.” Like hypotheses, such assertions can be shown to be correct or incorrect. A statement of fact or a hypothesis is a positive statement. Note also that positive statements can be false, but as long as they are testable, they are positive.

Normative Statements

Although people often disagree about positive statements, such disagreements can ultimately be resolved through investigation. There is another category of assertions, however, for which investigation can never resolve differences. A normative statement is one that makes a value judgment. Such a judgment is the opinion of the speaker; no one can “prove” that the statement is or is not correct. Here are some examples of normative statements in economics:

- We ought to do more to help the poor.
- People in the United States should save more for retirement.
- Corporate profits are too high.

These statements are based on the values of the person who makes them and can’t be proven false.

Because people have different values, normative statements often provoke disagreement. An economist whose values lead him or her to conclude that we should provide more help for the poor will disagree with one whose values lead to a conclusion that we should not. Because no test exists for these values, these two economists will continue to disagree, unless one persuades the other to adopt a different set of values. Many of the disagreements among economists are based on such differences in values and therefore are unlikely to be resolved.

It’s not uncommon for people to present an argument as positive, to make it more convincing to an audience, when in fact it has normative elements. Opinion pieces in newspapers or on other media are good examples of this. That’s why it’s important to be able to differentiate between positive and normative claims.

Try It

Visit this page in your course online to view this presentation.
PUTTING IT TOGETHER: CHOICE IN A WORLD OF SCARCITY

The Challenging Budget Constraints of a Student

We began this module with a discussion of the annual salaries of full-time U.S. workers with different levels of education. Let’s return to the very real economic issues that face most students when making decisions about their education.

First, we discussed the cost of choices and trade-offs and used the budget constraint model to demonstrate those costs. Each term, students make a trade-off between taking more credits in school and buying necessary items. Let’s create a budget constraint model for Camila, a community college student who is struggling to cover the cost of education. First, let’s assume that each credit hour costs $75. Camila wants to take 12 to 16 credits but also needs to pay for gas to drive between school, work, and other family responsibilities. Gas costs $3 per gallon. If she has a budget during the course of the academic term that allows her to spend a total of $1,500 on course credits and gas, what are Camila’s options?

Exercises

We can use the budget constraint equation to answer this question. 

**Step 1.** Apply the budget constraint equation to the scenario.

In Camila’s case, this works out to be

$$\text{Budget} = P_1 \times Q_1 + P_2 \times Q_2$$
Budget = 1500

\[ P_1 = 3 \text{ (price for a gallon of gas)} \]
\[ Q_1 = \text{gallons of gas (variable)} \]
\[ P_2 = 75 \text{ (price per credit hour)} \]
\[ Q_2 = \text{number of credit hours (variable)} \]

For Camila, this is
\[ 1500 = 3 \times Q_1 + 75 \times Q_2 \]

**Step 2.** Simplify the equation.

At this point we need to decide whether to solve for \( Q_1 \) or \( Q_2 \).

Remember, Camila was hoping to take at least 12 credit hours, so we know the value for \( Q_2 \). We will solve for \( Q_1 \) because, in this equation, it represents the number of gallons of gas Camila can pay for, depending on how many credit hours she takes during the academic term.

We are going solve for \( Q_1 \). First we will write the equation with the variables on the left to make solving easier:
\[ 3Q_1 + 75Q_2 = 1500. \]

\[ 3Q_1 + 75Q_2 = 1500 \]

\[ 3Q_1 = 1500 - 75Q_2 \quad \text{isolate } Q_1 \text{ on one side} \]

\[ \frac{3Q_1}{3} = \frac{1500}{3} - \frac{75Q_2}{3} \quad \text{divide everything by 3} \]

\[ Q_1 = 500 - 25Q_2 \]

**Step 3.** Use the equation.

We know that Camila hopes to take 12 credit hours during a term. \( Q_1 \) represents the number of credits she hopes to fund, so we plug in 12 for \( Q_2 \), which gives us
\[ Q_1 = 500 - 25(12) \]
\[ Q_1 = 500 - 300 \]
\[ Q_1 = 200 \]

This means that Camila can buy 200 gallons of gas during the term she is taking 12 credit hours (point M on the graph, below).

If you plug other numbers of credit hours into the equation, you get the results shown in Table 1, below.

<table>
<thead>
<tr>
<th>Table 1. Camila’s Budget Constraint</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Point</strong></td>
</tr>
<tr>
<td>A</td>
</tr>
<tr>
<td>B</td>
</tr>
<tr>
<td>C</td>
</tr>
<tr>
<td>D</td>
</tr>
<tr>
<td>E</td>
</tr>
</tbody>
</table>
Table 1. Camila’s Budget Constraint

<table>
<thead>
<tr>
<th>Point</th>
<th>Number of Credit Hours</th>
<th>Gallons of Gas</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>5</td>
<td>375</td>
</tr>
<tr>
<td>G</td>
<td>6</td>
<td>350</td>
</tr>
<tr>
<td>H</td>
<td>7</td>
<td>325</td>
</tr>
<tr>
<td>I</td>
<td>8</td>
<td>300</td>
</tr>
<tr>
<td>J</td>
<td>9</td>
<td>275</td>
</tr>
<tr>
<td>K</td>
<td>10</td>
<td>250</td>
</tr>
<tr>
<td>L</td>
<td>11</td>
<td>225</td>
</tr>
<tr>
<td>M</td>
<td>12</td>
<td>200</td>
</tr>
<tr>
<td>N</td>
<td>13</td>
<td>175</td>
</tr>
<tr>
<td>O</td>
<td>14</td>
<td>150</td>
</tr>
<tr>
<td>P</td>
<td>15</td>
<td>125</td>
</tr>
<tr>
<td>Q</td>
<td>16</td>
<td>100</td>
</tr>
</tbody>
</table>

**Step 4.** Graph the results.

If we plot each point on a graph, as below, we can see a line that shows us the number of credit hours that Camila can fund while still paying for gas.

*Figure 1. Camila’s Budget Constraint.*
Education and the Production Possibilities Curve

As state legislators allocate funding, they often make independent decision about the funding amount and approach for education and the funding amount and approach for corrections (or prisons). Economists recognize that these are not independent decisions. The production possibilities curve demonstrates that if society invests more in prisons, there are will be a reduction in the resources available to invest in education.

The graph above demonstrates the trade-off between devoting resources to corrections and to education. If the society were to allocate all of its resources to corrections, it could produce at point A, but it would not have any resources to produce education. If it were to allocate all of its resources to education, it could produce at point F. Alternatively, society could choose to produce any combination of corrections and education shown on the production possibilities frontier.

Sometimes legislators don’t recognize the direct trade-off between investing in education and investing in prisons, but inevitably economists will point out the connection, and the press will jump in and question the legislators’ decision. In a world of scarcity, more spending in one necessarily means less to spend in others.

Should Society Invest in Prisons or Education?

As we consider the trade-offs between investments in prisons and education, is there a definitive “right” answer? Consider the following analysis by the Center on Budget and Policy Priorities:

Even as states spend more on corrections, they are underinvesting in educating children and young adults, especially those in high-poverty neighborhoods. At least 30 states are providing less general funding per student this year for K–12 schools than before the recession, after adjusting for inflation; in 14 states the reduction exceeds 10 percent. Higher education cuts have been even deeper: the average state has cut higher education funding per student by 23 percent since the recession hit, after adjusting for inflation. Eleven states spent more of their general funds on corrections than on higher education in 2013. And some of the states with the biggest education cuts in recent years also have among the nation’s highest incarceration rates.

This is not sound policy. State economies would be much stronger over time if states invested more in education and other areas that can boost long-term economic growth and less in maintaining extremely high prison populations. The economic health of many low-income neighborhoods, which face disproportionately high incarceration rates, could particularly improve if states reordered their spending in such a way. States could use the freed-up funds in a number of ways, such as expanding access to high-quality preschool, reducing class sizes in high-poverty schools, and revising state funding formulas to invest more in high-poverty neighborhoods. (Note: http://www.cbpp.org/sites/default/files/atoms/files/10-28-14sfp.pdf)
While the analysis cited is thorough and logical, the report above includes a range of positive and normative statements. If you reread the analysis with that in mind, you will find examples of both.

Positive Statements

- At least 30 states are providing less general funding per student this year for K–12 schools than before the recession, after adjusting for inflation; in 14 states the reduction exceeds 10 percent.
- Higher education cuts have been even deeper: the average state has cut higher education funding per student by 23 percent since the recession hit, after adjusting for inflation.
- Eleven states spent more of their general funds on corrections than on higher education in 2013. And some of the states with the biggest education cuts in recent years also have among the nation’s highest incarceration rates.

Normative Statements

- Even as states spend more on corrections, they are underinvesting in educating children and young adults, especially those in high-poverty neighborhoods.
- This is not sound policy.
- State economies would be much stronger over time if states invested more in education and other areas that can boost long-term economic growth and less in maintaining extremely high prison populations.

As you can see, your experience as a student affords you an important view into the trade-offs that are core to economics.

License & Attributions

CC licensed content, Original

- Putting It Together: Choice in a World of Scarcity. Provided by: Lumen Learning. License: CC BY Attribution
MODULE 3: SUPPLY AND DEMAND

WHY IT MATTERS: SUPPLY AND DEMAND

Why think about supply and demand?

Do you pay attention to the cost of a cup of coffee? Most people recognize that when they make coffee at home it's cheaper than buying a cup of coffee that someone else has made. You've probably also noticed that some coffee places are more expensive than others—a cup of coffee at Starbucks usually costs more than one at a gas station, for instance. Regardless of where you decide to buy coffee, the price can change dramatically.

![Figure 1. Coffee Prices. Source: Trading Economics](image)

While retailers make decisions about how much they will mark up the coffee drinks they sell, the underlying coffee prices all around the world are driven by supply and demand. Brazil accounts for 33 percent of all coffee production in the world, and in 2011 the country experienced a drought. Coffee trees had already been weakened by environmental factors, and the drought had a significant impact on production levels.

How did individuals react to this kind of shortage? Did the shortage have an impact on price? Take a look at the graph in Figure 1, above. You can see from the sharp spike in 2011 that a shortage in the supply of coffee did indeed have an impact on price. But what do we know about the demand for coffee during that time? Did coffee consumption levels fall? To answer these questions, we need to know more about how buyers and sellers interact in the marketplace. In short, we need to understand supply and demand.

In this section you’ll learn about these key economic factors and the laws that govern them. Understanding supply and demand is not only essential to the study of economics—it may also help you be a better-informed consumer and make
knowledgeable decisions about everything from your next cup of joe to your next job.

INTRODUCTION TO ECONOMIC SYSTEMS

What you’ll learn to do: describe and differentiate between major economic systems

Think about what a complex system a modern economy is. It includes all production of goods and services, all buying and selling, all employment. The economic life of every individual is interrelated, at least to a small extent, with the economic lives of thousands or even millions of other individuals. Who organizes and coordinates this system? Who insures that, for example, the number of televisions a society produces is the same as the amount it needs and wants? Who insures that the right number of employees works in the electronics industry? Who insures that televisions are produced in the best way possible? How does it all get done?

The answer to these important questions depends on the kind of economic system a society uses.

In this section, you’ll learn about the basic organizing principles of different types of economies. Understanding the characteristics of a competitive market, in particular, is an important foundation for understanding the mechanisms of supply and demand.

ECONOMIC SYSTEMS
Learning Objectives

- Describe characteristics of market economies, including free and competitive markets
- Describe characteristics of a planned, or command, economy

Figure 1. Perhaps a picture of a planned economy?

Types of Economies

In the modern world today, there is a range of economic systems, from market economies to planned (or command) economies.

Market Economies

A market is any situation that brings together buyers and sellers of goods or services. Buyers and sellers can be either individuals or businesses. In a market economy, economic decision-making happens through markets. Market economies are based on private enterprise: the means of production (resources and businesses) are owned and operated by private individuals or groups of private individuals. Businesses supply goods and services based on demand. Which goods and services are supplied depends on what products businesses think will bring them the most profit. The more a product is demanded by consumers or other businesses, the higher the price businesses can charge, and so the more of the product will be supplied. Consumer demand depends on peoples’ incomes. A person’s income is based on his or her ownership of resources (especially labor). The more society values the person’s output, the higher the income they will earn (think Lady Gaga or LeBron James).

Examples of free-market economies include Hong Kong, Singapore, and to a large extent, New Zealand, and the United States.

Free Markets

In a market economy, decisions about what products are available and at what prices are determined through the interaction of supply and demand. A competitive market is one in which there is a large number of buyers and sellers, so that no one can control the market price. A free market is one in which the government does not intervene in any
way. A free and competitive market economy is the ideal type of market economy, because what is supplied is exactly what consumers demand.

Price controls are an example of a market that is not free. When government intervenes, the market outcomes will be different from those that would occur in a free and competitive market model. When markets are less than perfectly competitive (e.g., monopolistic), the market outcomes will also differ.

Try It
Visit this page in your course online to view this presentation.

Planned (or Command) Economies

Command economies operate very differently. In a command economy, economic effort is devoted to goals passed down from a ruler or ruling class. Ancient Egypt was a good example: a large part of economic life was devoted to building pyramids (like the one at the left), for the pharaohs. Medieval manor life is another example: The lord provided the land for growing crops and protection in the event of war. In return, vassals provided labor and soldiers to do the lord's bidding. In the last century, communist countries have employed command economies.

In a command economy, resources and businesses are owned by the government. The government decides what goods and services will be produced and what prices will be charged for them. The government decides what methods of production will be used and how much workers will be paid. Some necessities like health care and education are provided for free, as long as the state determines that you need them. With the collapse of the former Soviet Union in the 1990s, command economies fell out of favor as an economic system. Currently, only North Korea and Cuba have command economies.

The primary distinction between a free and command economy is the degree to which the government determines what can be produced and what prices will be charged. In a free market, these determinations are made by the collective decisions of the market itself (which is comprised of producers and consumers). Producers and consumers make rational decisions about what will satisfy their self-interest and maximize profits, and the market responds accordingly. In a planned economy, the government makes most decisions about what will be produced and what the prices will be, and consumers react passively to that plan.

Most economies in the real world are mixed; they combine elements of command and market systems. The U.S. economy is positioned toward the market-oriented end of the spectrum. Many countries in Europe and Latin America, while primarily market-oriented, have a greater degree of government involvement in economic decisions than in the U.S. economy. China and Russia, while they are closer now to having a market-oriented system than several decades ago, remain closer to the command-economy end of the spectrum.

Try It
Visit this page in your course online to view this presentation.

Watch It

The following Crash Course video provides additional information about the broad economic choices that countries make when they decide between planned and market economies. The narrators talk fast, so you'll need to listen closely and possibly watch the video a second time!

Watch this video online: https://youtu.be/B43YEW2FvDs
Economic systems determine the following:

- What to produce?
- How to produce it?
- Who gets it?

In a planned economy, government controls the factors of production:

- In a true communist economy, there is no private property—everyone owns the factors of production. This type of planned economy is called a **command economy**
- In a socialist economy, there is some private property and some private control of industry.

In a free-market (capitalist) economy, individuals own the factors of production:

- Businesses produce products.
- Consumers choose the products they prefer leading the companies that produce them to make more profit.

Even in free markets, governments should

- Maintain the rule of law
- Create public goods and services such as roads and education
- Step in when the market gets things wrong (e.g., setting minimum wage, establishing environmental standards)

---

**Glossary**

**command economy**: an economy where economic decisions are passed down from government authority and where resources are owned by the government

**competitive market**: is one in which there is a large number of buyers and sellers, so that no one can control the market price

**free market**: a market in which the government does not intervene in any way

**market economy**: an economy where economic decisions are decentralized, resources are owned by private individuals, and businesses supply goods and services based on demand

---

**INTRODUCTION TO DEMAND**

What you’ll learn to do: explain the determinants of demand
Imagine that Ben & Jerry’s has a promotion to discount the price of their ice cream next summer. What do you think will happen to the amount of Ben & Jerry’s ice cream that people will want to buy? The likely result is that people will buy more ice cream. By the same token, if the price of the ice cream were to rise next summer, then ice cream purchases would likely fall. In this section, you will examine this “law of demand” and see why this simple concept is essential to understanding economics.

WHAT IS DEMAND?

Learning Objectives

- Explain demand and the law of demand
- Identify and explain a demand curve
- Create and interpret a demand curve using a data set
Demand for Goods and Services

Economists use the term demand to refer to the amount of some good or service consumers are willing and able to purchase at each price. Demand is based on needs and wants—a consumer may be able to differentiate between a need and a want, but from an economist's perspective, they are the same thing. Demand is also based on ability to pay. If you can't pay for it, you have no effective demand.

What a buyer pays for a unit of the specific good or service is called the price. The total number of units purchased at that price is called the quantity demanded. A rise in the price of a good or service almost always decreases the quantity of that good or service demanded. Conversely, a fall in price will increase the quantity demanded. When the price of a gallon of gasoline goes up, for example, people look for ways to reduce their consumption by combining several errands, commuting by carpool or mass transit, or taking weekend or vacation trips closer to home. Economists call this inverse relationship between price and quantity demanded the law of demand. The law of demand assumes that all other variables that affect demand are held constant.

An example from the market for gasoline can be shown in the form of a table or a graph. A table that shows the quantity demanded at each price, such as Table 1, is called a demand schedule. Price in this case is measured in dollars per gallon of gasoline. The quantity demanded is measured in millions of gallons over some time period (for example, per day or per year) and over some geographic area (like a state or a country).

<table>
<thead>
<tr>
<th>Price (per gallon)</th>
<th>Quantity Demanded (millions of gallons)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$1.00</td>
<td>800</td>
</tr>
<tr>
<td>$1.20</td>
<td>700</td>
</tr>
<tr>
<td>$1.40</td>
<td>600</td>
</tr>
<tr>
<td>$1.60</td>
<td>550</td>
</tr>
<tr>
<td>$1.80</td>
<td>500</td>
</tr>
</tbody>
</table>
Table 1. Price and Quantity Demanded of Gasoline

<table>
<thead>
<tr>
<th>Price (per gallon)</th>
<th>Quantity Demanded (millions of gallons)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$2.00</td>
<td>460</td>
</tr>
<tr>
<td>$2.20</td>
<td>420</td>
</tr>
</tbody>
</table>

A demand curve shows the relationship between price and quantity demanded on a graph like Figure 2, below, with price per gallon on the vertical axis and quantity on the horizontal axis. Note that this is an exception to the normal rule in mathematics that the independent variable (x) goes on the horizontal axis and the dependent variable (y) goes on the vertical. Economics is different from math! Note also that each point on the demand curve comes from one row in Table 1. For example, the upper most point on the demand curve corresponds to the last row in Table 1, while the lower most point corresponds to the first row.

![Demand Curve for Gasoline](https://example.com/demand_curve.png)

Figure 2. A Demand Curve for Gasoline (derived from the data in Table 1).

The demand schedule (Table 1) shows that as price rises, quantity demanded decreases, and vice versa. These points can then be graphed, and the line connecting them is the demand curve (shown by line D in the graph, above). The downward slope of the demand curve again illustrates the law of demand—the inverse relationship between prices and quantity demanded.

The demand schedule shown by Table 1 and the demand curve shown by the graph in Figure 2 are two ways of describing the same relationship between price and quantity demanded.

Try It

Visit this page in your course online to view this presentation.

Watch It

The demand curve shows how much of a good people are willing to buy at different prices. Watch this video to see an example of the demand for oil. When oil prices are high, fewer people are willing to pay the hefty price tag but some consumers, like airliners, depend so heavily on using oil for fuel, they are willing to pay a lot. Other low-value consumers will be less likely to pay for expensive oil, as they could find substitutes or alternatives.

Watch this video online: [https://youtu.be/kUPm2tMcbGE](https://youtu.be/kUPm2tMcbGE)

Demand curves will look somewhat different for each product. They may appear relatively steep or flat, or they may be straight or curved. Nearly all demand curves share the fundamental similarity that they slope down from left to right. In
this way, demand curves embody the law of demand: As the price increases, the quantity demanded decreases, and conversely, as the price decreases, the quantity demanded increases.

Try It

Visit this page in your course online to view this presentation.

Demand vs. Quantity Demanded

In economic terminology, demand is not the same as quantity demanded. When economists talk about demand, they mean the relationship between a range of prices and the quantities demanded at those prices, as illustrated by a demand curve or a demand schedule. When economists talk about quantity demanded, they mean only a certain point on the demand curve, or one quantity on the demand schedule. In short, demand refers to the curve and quantity demanded refers to the (specific) point on the curve.

What Factors Affect Demand?

We defined demand as the amount of some product that a consumer is willing and able to purchase at each price. This suggests at least two factors, in addition to price, that affect demand. “Willingness to purchase” suggests a desire to buy, and it depends on what economists call tastes and preferences. If you neither need nor want something, you won’t be willing to buy it. “Ability to purchase” suggests that income is important. Professors are usually able to afford better housing and transportation than students, because they have more income. The prices of related goods can also affect demand. If you need a new car, for example, the price of a Honda may affect your demand for a Ford. Finally, the size or composition of the population can affect demand. The more children a family has, the greater their demand for car insurance and the less for diapers and baby formula.

These factors matter both for demand by an individual and demand by the market as a whole. Exactly how do these various factors affect demand, and how do we show the effects graphically? To answer those questions, we need the ceteris paribus assumption.

The Ceteris Paribus Assumption

A demand curve or a supply curve (which we’ll cover later in this module) is a relationship between two, and only two, variables: price on the vertical axis and quantity on the horizontal axis. The assumption behind a demand curve or a supply curve is that no relevant economic factors, other than the product’s price, are changing. Economists call this assumption ceteris paribus, a Latin phrase meaning “other things being equal.” Any given demand or supply curve is based on the ceteris paribus assumption that all else is held equal. Therefore, a demand curve or a supply curve is a relationship between two, and only two, variables when all other variables are held equal. If all else is not held equal, then the laws of supply and demand will not necessarily hold.

Try It

Visit this page in your course online to view this presentation.

WHEN DOES CETERIS PARIBUS APPLY?

Ceteris paribus is applied when we look at how changes in price affect demand or supply, but ceteris paribus can also be applied more generally. In the real world, demand and supply depend on more factors than just price. For example, a consumer’s demand depends on income, and a producer’s supply depends on the cost of producing the
product. How can we analyze the effect on demand or supply if multiple factors are changing at the same time—say price rises and income falls? The answer is that we examine the changes one at a time, and assume that the other factors are held constant.

For example, we can say that an increase in the price reduces the amount consumers will buy (assuming income, and anything else that affects demand, is unchanged). Additionally, a decrease in income reduces the amount consumers can afford to buy (assuming price, and anything else that affects demand, is unchanged). This is what the *ceteris paribus* assumption really means. In this particular case, after we analyze each factor separately, we can combine the results. The amount consumers buy falls for two reasons: first because of the higher price and second because of the lower income.

### Watch It

Watch this video to review the theory of demand. Remember that, according to the law of demand and all other things being equal (*ceteris paribus*):

- the lower the price of a product, the more of it will be bought
- the higher the price of a product, the less of it will be bought

Watch this video online: [https://youtu.be/uXlZIn6W7Ew](https://youtu.be/uXlZIn6W7Ew)

### Glossary

**ceteris paribus**: When changing one variable in a function (e.g. demand for some product), we assume everything else held constant

**demand**: the relationship between the price of a certain good or service and the quantity of that good or service someone is willing and able to buy

**demand curve**: a graphic representation of the relationship between price and quantity demanded of a certain good or service, with price on the vertical axis and quantity on the horizontal axis

**demand schedule**: a table that shows the quantity demanded for a certain good or service at a range of prices

**law of demand**: the common relationship that a higher price leads to a lower quantity demanded of a certain good or service and a lower price leads to a higher quantity demanded, while all other variables are held constant

**price**: what a buyer pays for a unit of the specific good or service

**quantity demanded**: the total number of units of a good or service consumers wish to purchase at a given price
FACTORS AFFECTING DEMAND

Learning Objectives

- Describe which factors cause a shift in the demand curve and explain why the shift occurs
- Define and give examples of substitutes and complements
- Draw a demand curve and graphically represent changes in demand

We know that a change in prices affects the quantity demanded. Price, however, is not the only thing that influences demand. For example, how is demand for vegetarian food affected if, say, health concerns cause more consumers to avoid eating meat?

The Effect of Income on Demand

Let’s use income as an example of how factors other than price affect demand. Figure 1 shows the initial demand for automobiles as D₀. At point Q, for example, if the price is $20,000 per car, the quantity of cars demanded is 18 million. D₀ also shows how the quantity of cars demanded would change as a result of a higher or lower price. For example, if the price of a car rose to $22,000, the quantity demanded would decrease to 17 million, at point R.

![Figure 1. Shifts in Demand: A Car Example.](image)

The original demand curve D₀, like every demand curve, is based on the ceteris paribus assumption that no other economically relevant factors change. Now imagine that the economy expands in a way that raises the incomes of many people, making cars more affordable. How will this affect demand? How can we show this graphically?

Return to Figure 1. The price of cars is still $20,000, but with higher incomes, the quantity demanded has now increased to 20 million cars, shown at point S. As a result of the higher income levels, the demand curve shifts to the right to the new demand curve D₁, indicating an increase in demand. Table 1, below, shows clearly that this increased demand would occur at every price, not just the original one.

<table>
<thead>
<tr>
<th>Price</th>
<th>Decrease to D₂</th>
<th>Original Quantity Demanded D₀</th>
<th>Increase to D₁</th>
</tr>
</thead>
</table>

Table 1. Price and Demand Shifts: A Car Example
Table:

<table>
<thead>
<tr>
<th>Price</th>
<th>Decrease to $D_2$</th>
<th>Original Quantity Demanded $D_0$</th>
<th>Increase to $D_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$16,000</td>
<td>17.6 million</td>
<td>22.0 million</td>
<td>24.0 million</td>
</tr>
<tr>
<td>$18,000</td>
<td>16.0 million</td>
<td>20.0 million</td>
<td>22.0 million</td>
</tr>
<tr>
<td>$20,000</td>
<td>14.4 million</td>
<td>18.0 million</td>
<td>20.0 million</td>
</tr>
<tr>
<td>$22,000</td>
<td>13.6 million</td>
<td>17.0 million</td>
<td>19.0 million</td>
</tr>
<tr>
<td>$24,000</td>
<td>13.2 million</td>
<td>16.5 million</td>
<td>18.5 million</td>
</tr>
<tr>
<td>$26,000</td>
<td>12.8 million</td>
<td>16.0 million</td>
<td>18.0 million</td>
</tr>
</tbody>
</table>

Now, imagine that the economy slows down so that many people lose their jobs or work fewer hours, reducing their incomes. In this case, the decrease in income would lead to a lower quantity of cars demanded at every given price, and the original demand curve $D_0$ would shift left to $D_2$. The shift from $D_0$ to $D_2$ represents such a decrease in demand: At any given price level, the quantity demanded is now lower. In this example, a price of $20,000 means 18 million cars sold along the original demand curve, but only 14.4 million sold after demand fell.

When a demand curve shifts, it does not mean that the quantity demanded by every individual buyer changes by the same amount. In this example, not everyone would have higher or lower income and not everyone would buy or not buy an additional car. Instead, a shift in a demand curve captures a pattern for the market as a whole: Increased demand means that at every given price, the quantity demanded is higher, so that the demand curve shifts to the right from $D_0$ to $D_1$. And, decreased demand means that at every given price, the quantity demanded is lower, so that the demand curve shifts to the left from $D_0$ to $D_2$.

We just argued that higher income causes greater demand at every price. This is true for most goods and services. For some—luxury cars, vacations in Europe, and fine jewelry—the effect of a rise in income can be especially pronounced. A product whose demand rises when income rises, and vice versa, is called a normal good. A few exceptions to this pattern do exist, however. As incomes rise, many people will buy fewer generic-brand groceries and more name-brand groceries. They are less likely to buy used cars and more likely to buy new cars. They will be less likely to rent an apartment and more likely to own a home, and so on. A product whose demand falls when income rises, and vice versa, is called an inferior good. In other words, when income increases, the demand curve shifts to the left.

Try It

Visit this page in your course online to view this presentation.

Watch It

A change in price does not shift the demand curve. It only shows a difference in the quantity demanded. The demand curve will move left or right when there is an underlying change in demand at all prices.

Watch this video online: https://youtu.be/aTSwcXJ700c

Other Factors That Shift Demand Curves
Income is not the only factor that causes a shift in demand. Other things that change demand include tastes and preferences, the composition or size of the population, the prices of related goods, and even expectations. A change in any one of the underlying factors that determine what quantity people are willing to buy at a given price will cause a shift in demand. Graphically, the new demand curve lies either to the right (an increase) or to the left (a decrease) of the original demand curve. Let’s look at these factors.

Changing Tastes or Preferences

From 1980 to 2012, the per-person consumption of chicken by Americans rose from 33 pounds per year to 81 pounds per year, and consumption of beef fell from 77 pounds per year to 57 pounds per year, according to the U.S. Department of Agriculture (USDA). Changes like these are largely due to shifts in taste, which change the quantity of a good demanded at every price: That is, they shift the demand curve for that good—rightward for chicken and leftward for beef.

Changes in the Composition of the Population

The proportion of elderly citizens in the United States population is rising. It rose from 9.8 percent in 1970 to 12.6 percent in 2000 and will be a projected (by the U.S. Census Bureau) 20 percent of the population by 2030. A society with relatively more children, like the United States in the 1960s, will have greater demand for goods and services like tricycles and day care facilities. A society with relatively more elderly persons, as the United States is projected to have by 2030, has a higher demand for nursing homes and hearing aids. Similarly, changes in the size of the population can affect the demand for housing and many other goods. Each of these changes in demand will be shown as a shift in the demand curve.

Changes in the Prices of Related Goods

The demand for a product can also be affected by changes in the prices of related goods such as substitutes or complements. A substitute is a good or service that can be used in place of another good or service. As electronic books become more available, you would expect to see a decrease in demand for traditional printed books. A lower price for a substitute decreases demand for the other product. For example, in recent years as the price of tablet computers has fallen, the quantity demanded has increased (because of the law of demand). Since people are purchasing tablets, there has been a decrease in demand for laptops, which can be shown graphically as a leftward shift in the demand curve for laptops. A higher price for a substitute good has the reverse effect.

Other goods are complements for each other, meaning that the goods are often used together, because consumption of one good tends to enhance consumption of the other. Examples include breakfast cereal and milk; notebooks and pens or pencils, golf balls and golf clubs; gasoline and sport utility vehicles; and the five-way combination of bacon, lettuce, tomato, mayonnaise, and bread. If the price of golf clubs rises, since the quantity of golf clubs demanded falls (because of the law of demand), demand for a complement good like golf balls decreases, too. Similarly, a higher price for skis would shift the demand curve for a complement good like ski resort trips to the left, while a lower price for a complement has the reverse effect.

Changes in Expectations About Future Prices

While it is clear that the price of a good affects the quantity demanded, it is also true that expectations about the future price (or expectations about tastes and preferences, income, and so on) can affect demand. For example, if people hear that a hurricane is coming, they may rush to the store to buy flashlight batteries and bottled water. If people learn that the price of a good like coffee is likely to rise in the future, they may head for the store to stock up on coffee now. These
changes in demand are shown as shifts in the curve. Therefore, a **shift in demand** happens when a change in some economic factor (other than the current price) causes a different quantity to be demanded at every price.

**Try It**

Visit this page in your course online to view this presentation.

---

**Exercise: Shift in Demand Due to Income Increase**

A shift in demand means that at any price (and at every price), the quantity demanded will be different than it was before. Following is a graphic illustration of a shift in demand due to an income increase.

**Step 1.** Draw the graph of a demand curve for a normal good like pizza. Pick a price (like $P_0$). Identify the corresponding $Q_0$. An example is shown in Figure 5.

![Figure 5. Demand Curve. A demand curve can be used to identify how much consumers would buy at any given price.](image)

**Step 2.** Suppose income increases. As a result of the change, are consumers going to buy more or less pizza? The answer is more. Draw a dotted horizontal line from the chosen price, through the original quantity demanded, to the new point with the new $Q_1$. Draw a dotted vertical line down to the horizontal axis and label the new $Q_1$. An example is provided in Figure 6.

![Figure 6. Demand Curve with Income Increase. With an increase in income, consumers will purchase larger quantities, pushing demand to the right.](image)

**Step 3.** Now, shift the curve through the new point. You will see that an increase in income causes an upward (or rightward) shift in the demand curve, so that at any price, the quantities demanded will be higher, as shown in Figure 7.
Figure 7. Demand Curve Shifted Right. With an increase in income, consumers will purchase larger quantities, pushing demand to the right, and causing the demand curve to shift right.

Figure 8. Remember that changes in price change the point of quantity demanded on the demand curve, but changes in other factors (such as taste, population, income, expectations, and prices of other goods) will cause the entire demand curve to shift.

Summary: What Factors Shift Demand?

Six factors that can shift demand curves are summarized in Figure 9, below. The direction of the arrows indicates whether the demand curve shifts represent an increase in demand or a decrease in demand. Notice that a change in the price of the good or service itself is not listed among the factors that can shift a demand curve. A change in the price of a good or service causes a movement along a specific demand curve, and it typically leads to some change in the quantity demanded, but it does not shift the demand curve.
Figure 9. Factors That Shift Demand Curves (a) A list of factors that can cause an increase in demand from $D_0$ to $D_1$. (b) The same factors, if their direction is reversed, can cause a decrease in demand from $D_0$ to $D_1$.

Try It

Visit this page in your course online to view this presentation.

Glossary

complements: goods or services that are used together because the use of one enhances the use of the other

substitutes: goods or services that can be used in place of one another

inferior good: good or service whose demand decreases when a consumer’s income increases and demand increases when income decreases

normal good: good or service whose demand increases when a consumer’s income increases and demand decreases when income decreases

LEARN BY DOING: GRAPHING DEMAND

Try It

These questions allow you to get as much practice as you need, as you can click the link at the top of the first question (“Try another version of these questions”) to get a new set of questions. Practice until you feel comfortable
LEARN BY DOING: DEMAND FOR FOOD TRUCKS

Try It

Play the simulation below to examine how the demand for drinks changes when the price or the weather change. The simulation allow unlimited attempts so that you can gain experience applying the concepts. Visit this page in your course online to view this presentation.

INTRODUCTION TO SUPPLY

What you’ll learn to do: explain the determinants of supply
So far you've learned about the role of demand in economics—which is the consumer side of the story. In this section, you'll learn about the producer side of economics to see what factors impact the amount of goods supplied in a market. For example, suppose the global price of petroleum falls significantly. What do you think will happen to the supply of gasoline? How are supply and price connected? In this section you'll examine the law of supply and see why this counterpart to “demand” is also essential to understanding economics.

WHAT IS SUPPLY?

Learning Objectives

- Explain supply and the law of supply
- Identify and explain a supply curve
- Create and interpret a supply curve using a data set

Supply of Goods and Services

When economists talk about supply, they mean the amount of some good or service a producer is willing to supply at each price. Price is what the producer receives for selling one unit of a good or service. A rise in price almost always leads to an increase in the quantity supplied of that good or service, while a fall in price will decrease the quantity supplied. When the price of gasoline rises, for example, it encourages profit-seeking firms to take several actions: expand exploration for oil reserves; drill for more oil; invest in more pipelines and oil tankers to bring the oil to plants where it can be refined into gasoline; build new oil refineries; purchase additional pipelines and trucks to ship the gasoline to gas stations; and open more gas stations or keep existing gas stations open longer hours. Economists call this positive relationship between price and quantity supplied—that a higher price leads to a higher quantity supplied and a lower price leads to a lower quantity supplied—the law of supply. The law of supply, like the law of demand, assumes that all other variables that affect supply are held equal (ceteris paribus).

Try It

Visit this page in your course online to view this presentation.

Watch It
The supply curve shows how much that sellers will be willing to provide at different prices. Because suppliers want to make a profit, companies have an incentive to sell more oil if it sells at a higher price.

Watch this video online: https://youtu.be/nKvrbOq1Ofl

Is Supply the same as Quantity Supplied?

In economic terminology, supply is not the same as quantity supplied. When economists refer to supply, they mean the relationship between a range of prices and the quantities supplied at those prices, a relationship that can be illustrated with a supply curve or a supply schedule. When economists refer to quantity supplied, they mean only a certain point on the supply curve, or one quantity on the supply schedule. In short, supply refers to the curve, and quantity supplied refers to the (specific) point on the curve.

Like demand, supply can be illustrated using a table or a graph. A supply schedule is a table—like Table 1, below—that shows the quantity supplied at a range of different prices. Again, price is measured in dollars per gallon of gasoline, and quantity supplied is measured in millions of gallons.

<table>
<thead>
<tr>
<th>Price (per gallon)</th>
<th>Quantity Supplied (millions of gallons)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$1.00</td>
<td>500</td>
</tr>
<tr>
<td>$1.20</td>
<td>550</td>
</tr>
<tr>
<td>$1.40</td>
<td>600</td>
</tr>
<tr>
<td>$1.60</td>
<td>640</td>
</tr>
<tr>
<td>$1.80</td>
<td>680</td>
</tr>
<tr>
<td>$2.00</td>
<td>700</td>
</tr>
<tr>
<td>$2.20</td>
<td>720</td>
</tr>
</tbody>
</table>

A supply curve is a graphic illustration of the relationship between price, shown on the vertical axis, and quantity, shown on the horizontal axis. Figure 1 illustrates the law of supply, again using the market for gasoline as an example. You can see from this curve that as the price rises, quantity supplied also increases and vice versa. The supply schedule and the supply curve are just two different ways of showing the same information. Note that each point on the supply curve comes from one row in Table 1. For example, the lowermost point on the supply curve corresponds to the first row in Table 1, while the upper most point corresponds to the last row. Notice also that the horizontal and vertical axes on the graph for the supply curve are the same as for the demand curve.
The shape of supply curves will vary somewhat according to the product: steeper, flatter, straighter, or curved. Nearly all supply curves, however, share a basic similarity: They slope up from left to right and illustrate the law of supply. As the price rises, say, from $1.00 per gallon to $2.20 per gallon, the quantity supplied increases from 500 gallons to 720 gallons. Conversely, as the price falls, the quantity supplied decreases.

**glossary**

**law of supply**: the common relationship that a higher price leads to a higher quantity supplied of a certain good or service and a lower price leads to a higher quantity supplied, while all other variables are held constant

**quantity supplied**: the total number of units of a good or service producers are willing to supply at a given price

**supply**: the relationship between the price of a certain good or service and the quantity of that good or service producers are willing to offer for sale

**supply curve**: a graphic representation of the relationship between price and quantity supplied of a certain good or service, with price on the vertical axis and quantity on the horizontal axis

**supply schedule**: a table that shows the quantity demanded for a certain good or service at a range of prices
FACTORS AFFECTING SUPPLY

Learning Objectives

- Describe which factors cause a shift in the supply curve and show them on a graph

How Production Costs Affect Supply

A supply curve shows how quantity supplied will change as the price rises and falls, assuming *ceteris paribus*, that is, no other economically relevant factors are changing. If other factors relevant to supply do change, then the entire supply curve will shift. Just as a shift in demand is represented by a change in the quantity demanded at every price, a shift in supply means a change in the quantity supplied at every price.

In thinking about the factors that affect supply, remember what motivates firms: profits, which are the difference between revenues and costs. Goods and services are produced using combinations of labor, materials, and machinery, or what we call inputs (also called factors of production). If a firm faces lower costs of production, while the prices for the good or service the firm produces remain unchanged, a firm’s profits go up. When a firm’s profits increase, it’s more motivated to produce output (goods or services), since the more it produces the more profit it will earn. So, when costs of production fall, a firm will tend to supply a larger quantity at any given price for its output. This can be shown by the supply curve shifting to the right.

Take, for example, a messenger company that delivers packages around a city. The company may find that buying gasoline is one of its main costs. If the price of gasoline falls, then the company will find it can deliver packages more cheaply than before. Since lower costs correspond to higher profits, the messenger company may now supply more of its services at any given price. For example, given the lower gasoline prices, the company can now serve a greater area, and increase its supply.

Conversely, if a firm faces higher costs of production, then it will earn lower profits at any given selling price for its products. As a result, a higher cost of production typically causes a firm to supply a smaller quantity at any given price. In this case, the supply curve shifts to the left.

Consider the supply for cars, shown by curve $S_0$ in Figure 2, below. Point J indicates that if the price is $20,000, the quantity supplied will be 18 million cars. If the price rises to $22,000 per car, *ceteris paribus*, the quantity supplied will rise to 20 million cars, as point K on the $S_0$ curve shows. The same information can be shown in table form, as in Table 1.
Table 1. Price and Shifts in Supply: A Car Example

<table>
<thead>
<tr>
<th>Price</th>
<th>Decrease to $S_1$</th>
<th>Original Quantity Supplied $S_0$</th>
<th>Increase to $S_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$16,000$</td>
<td>10.5 million</td>
<td>12.0 million</td>
<td>13.2 million</td>
</tr>
<tr>
<td>$18,000$</td>
<td>13.5 million</td>
<td>15.0 million</td>
<td>16.5 million</td>
</tr>
<tr>
<td>$20,000$</td>
<td>16.5 million</td>
<td>18.0 million</td>
<td>19.8 million</td>
</tr>
<tr>
<td>$22,000$</td>
<td>18.5 million</td>
<td>20.0 million</td>
<td>22.0 million</td>
</tr>
<tr>
<td>$24,000$</td>
<td>19.5 million</td>
<td>21.0 million</td>
<td>23.1 million</td>
</tr>
</tbody>
</table>

Now imagine that the price of steel—an important component in vehicle manufacturing—rises, so that producing a car has become more expensive. At any given price for selling cars, car manufacturers will react by supplying a lower quantity. This can be shown graphically as a leftward shift of supply, from $S_0$ to $S_1$, which indicates that at any given price, the quantity supplied decreases. In this example, at a price of $20,000, the quantity supplied decreases from 18 million on the original supply curve ($S_0$) to 16.5 million on the supply curve $S_1$, which is labeled as point L.

Conversely, if the price of steel decreases, producing a car becomes less expensive. At any given price for selling cars, car manufacturers can now expect to earn higher profits, so they will supply a higher quantity. The shift of supply to the right, from $S_0$ to $S_2$, means that at all prices, the quantity supplied has increased. In this example, at a price of $20,000, the quantity supplied increases from 18 million on the original supply curve ($S_0$) to 19.8 million on the supply curve $S_2$, which is labeled M.

**Shift in Supply Due to Production-Cost Increase**

We know that a supply curve shows the minimum price a firm will accept to produce a given quantity of output. What happens to the supply curve when the cost of production goes up? Following is an example of a shift in supply due to an increase in production cost.
Exercise: Shift in Supply

**Step 1.** Draw a graph of a supply curve for pizza. Pick a quantity (like $Q_0$). If you draw a vertical line up from $Q_0$ to the supply curve, you will see the price the firm chooses. An example is shown in Figure 4.

![Figure 4. Supply Curve](image)

**Figure 4. Supply Curve.** The supply curve can be used to show the minimum price a firm will accept to produce a given quantity of output.

**Step 2.** Why did the firm choose that price and not some other? One way to think about this is that the price is composed of two parts. The first part is the average cost of production: in this case, the cost of the pizza ingredients (dough, sauce, cheese, pepperoni, and so on), the cost of the pizza oven, the rent on the shop, and the wages of the workers. The second part is the firm’s desired profit, which is determined, among other factors, by the profit margins in that particular business. If you add these two parts together, you get the price the firm wishes to charge. The quantity $Q_0$ and associated price $P_0$ give you one point on the firm’s supply curve, as shown in Figure 5.
Step 3. Now, suppose that the cost of production goes up. Perhaps cheese has become more expensive by $0.75 per pizza. If that is true, the firm will want to raise its price by the amount of the increase in cost ($0.75). Draw this point on the supply curve directly above the initial point on the curve, but $0.75 higher, as shown in Figure 6.

Step 4. Shift the supply curve through this point. You will see that an increase in cost causes a leftward shift of the supply curve so that at any price, the quantities supplied will be smaller, as shown in Figure 7.
In the example above, we saw that changes in the prices of inputs in the production process will affect the cost of production and thus the supply. Several other things affect the cost of production, too, such as changes in weather or other natural conditions, new technologies for production, and some government policies.

The cost of production for many agricultural products will be affected by changes in natural conditions. For example, the area of northern China that typically grows about 60 percent of the country’s wheat output experienced its worst drought in at least fifty years in the second half of 2009. A drought decreases the supply of agricultural products, which means that at any given price, a lower quantity will be supplied; conversely, especially good weather would shift the supply curve to the right.

When a firm discovers a new technology that allows it to produce at a lower cost, the supply curve will shift to the right, as well. For instance, in the 1960s a major scientific effort nicknamed the Green Revolution focused on breeding improved seeds for basic crops like wheat and rice. By the early 1990s, more than two-thirds of the wheat and rice in low-income countries around the world was grown with these Green Revolution seeds—and the harvest was twice as high per acre. A technological improvement that reduces costs of production will shift supply to the right, so that a greater quantity will be produced at any given price.

Government policies can affect the cost of production and the supply curve through taxes, regulations, and subsidies. For example, the U.S. government imposes a tax on alcoholic beverages that collects about $8 billion per year from producers. Taxes are treated as costs by businesses. Higher costs decrease supply for the reasons discussed above. Other examples of policy that can affect cost are the wide array of government regulations that require firms to spend money to provide a cleaner environment or a safer workplace; complying with regulations increases the cost of producing any level of output.

A government subsidy, on the other hand, is the opposite of a tax. A subsidy occurs when the government pays a firm directly or reduces the firm’s taxes if the firm carries out certain actions. From the firm’s perspective, subsidies are an offset to costs; they essentially reduce the cost of production and increase supply at every given price, shifting supply to the right.

Summary: What Factors Shift Supply?

Changes in the cost of inputs, natural disasters, new technologies, taxes, subsidies, and government regulation all affect the cost of production. In turn, these factors affect how much firms are willing to supply at any given price.

Figure 9 below summarizes factors that change the supply of goods and services. Notice that a change in the price of the product itself is not among the factors that shift the supply curve. Although a change in price of a good or service typically causes a change in quantity supplied or a movement along the supply curve for that specific good or service, it does not cause the supply curve itself to shift.

![Figure 8. Field of Wheat. Especially good growing seasons and weather could lead to greater supply and a rightward shift in the supply curve.](image)

![Figure 9. Factors That Shift Supply Curves.](image)
Because demand and supply curves appear on a two-dimensional diagram with only price and quantity on the axes, an unwary visitor to the land of economics might be fooled into believing that economics is about only four topics: demand, supply, price, and quantity. However, demand and supply are really “umbrella” concepts: demand covers all the factors that affect demand, and supply covers all the factors that affect supply. Factors other than price that affect demand and supply are included by using shifts in the demand or the supply curve. In this way, the two-dimensional demand and supply model becomes a powerful tool for analyzing a wide range of economic circumstances.

Try It

Visit this page in your course online to view this presentation.

Glossary

**subsidy**: a government payment to firms to encourage production of some good or service

---

LEARN BY DOING: GRAPHING SUPPLY AND DEMAND

Try It

These questions allow you to get as much practice as you need, as you can click the link at the top of the first question (“Try another version of these questions”) to get a new set of questions. Practice until you feel comfortable doing the questions and then move on. Note that you will use the information provided in the first question for all of the questions on this page. Visit this page in your course online to practice before taking the quiz.

---

LEARN BY DOING: SUPPLY OF FOOD TRUCKS
Try It

In this simulation, you will get to adjust how many food trucks are needed to feed some players and fans attending a soccer match. Click through the various options to see how different choices lead to different outcomes in the cost and amount of food supplied.
Visit this page in your course online to view this presentation.

INTRODUCTION TO EQUILIBRIUM

What you’ll learn to do: explain and graphically illustrate market equilibrium, surplus, and shortage

In this section, you’ll learn how supply and demand interact to determine the ideal price and quantity of a good in a market. When a good is not sold at its ideal price, a shortage or a surplus may be the result.

EQUILIBRIUM, SURPLUS, AND SHORTAGE
Learning Objectives

- Define equilibrium price and quantity and identify them in a market
- Define surpluses and shortages and explain how they cause the price to move towards equilibrium

Demand and Supply

In order to understand market equilibrium, we need to start with the laws of demand and supply. Recall that the law of demand says that as price decreases, consumers demand a higher quantity. Similarly, the law of supply says that when price decreases, producers supply a lower quantity.

Because the graphs for demand and supply curves both have price on the vertical axis and quantity on the horizontal axis, the demand curve and supply curve for a particular good or service can appear on the same graph. Together, demand and supply determine the price and the quantity that will be bought and sold in a market. These relationships are shown as the demand and supply curves in Figure 1, which is based on the data in Table 1, below.

![Figure 1. The supply and demand curves for gasoline.](image-url)

<table>
<thead>
<tr>
<th>Price (per gallon)</th>
<th>Quantity demanded (millions of gallons)</th>
<th>Quantity supplied (millions of gallons)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$1.00</td>
<td>800</td>
<td>500</td>
</tr>
<tr>
<td>$1.20</td>
<td>700</td>
<td>550</td>
</tr>
<tr>
<td>$1.40</td>
<td>600</td>
<td>600</td>
</tr>
<tr>
<td>$1.60</td>
<td>550</td>
<td>640</td>
</tr>
</tbody>
</table>
Table 1. Price, Quantity Demanded, and Quantity Supplied

<table>
<thead>
<tr>
<th>Price (per gallon)</th>
<th>Quantity demanded (millions of gallons)</th>
<th>Quantity supplied (millions of gallons)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$1.80</td>
<td>500</td>
<td>680</td>
</tr>
<tr>
<td>$2.00</td>
<td>460</td>
<td>700</td>
</tr>
<tr>
<td>$2.20</td>
<td>420</td>
<td>720</td>
</tr>
</tbody>
</table>

If you look at either Figure 1 or Table 1, you’ll see that at most prices the amount that consumers want to buy (which we call the quantity demanded) is different from the amount that producers want to sell (which we call the quantity supplied). What does it mean when the quantity demanded and the quantity supplied aren’t the same? The answer is: a surplus or a shortage.

Surplus or Excess Supply

Let’s consider one scenario in which the amount that producers want to sell doesn’t match the amount that consumers want to buy. Consider our gasoline market example. Imagine that the price of a gallon of gasoline were $1.80 per gallon. This price is illustrated by the dashed horizontal line at the price of $1.80 per gallon in Figure 2, below.

![Figure 2. A price above equilibrium creates a surplus.](image)

At this price, the quantity demanded is 500 gallons, and the quantity of gasoline supplied is 680 gallons. You can also find these numbers in Table 1, above. Now, compare the quantity demanded and quantity supplied at this price. Quantity supplied (680) is greater than quantity demanded (500). Or, to put it in words, the amount that producers want to sell is greater than the amount that consumers want to buy. We call this a situation of excess supply (since Qs > Qd) or a surplus. Note that whenever we compare supply and demand, it’s in the context of a specific price—in this case, $1.80 per gallon.

With a surplus, gasoline accumulates at gas stations, in tanker trucks, in pipelines, and at oil refineries. This accumulation puts pressure on gasoline sellers. If a surplus remains unsold, those firms involved in making and selling gasoline are not receiving enough cash to pay their workers and cover their expenses. In this situation, some firms will want to cut prices, because it is better to sell at a lower price than not to sell at all. Once some sellers start cutting prices; others will follow to avoid losing sales. These price reductions will, in turn, stimulate a higher quantity demanded.
How far will the price fall? Whenever there is a surplus, the price will drop until the surplus goes away. When the surplus is eliminated, the quantity supplied just equals the quantity demanded—that is, the amount that producers want to sell exactly equals the amount that consumers want to buy. We call this equilibrium, which means “balance.” In this case, the equilibrium occurs at a price of $1.40 per gallon and at a quantity of 600 gallons. You can see this in Figure 2 (and Figure 1) where the supply and demand curves cross. You can also find it in Table 1 (the numbers in bold).

Try It

Visit this page in your course online to view this presentation.

Shortage or Excess Demand

Let’s return to our gasoline problem. Suppose that the price is $1.20 per gallon, as the dashed horizontal line at this price in Figure 3, below, shows. At this price, the quantity demanded is 700 gallons, and the quantity supplied is 550 gallons.

![Figure 3. A price below equilibrium creates a shortage.](image)

Quantity supplied (550) is less than quantity demanded (700). Or, to put it in words, the amount that producers want to sell is less than the amount that consumers want to buy. We call this a situation of excess demand (since Qd > Qs) or a shortage.

In this situation, eager gasoline buyers mob the gas stations, only to find many stations running short of fuel. Oil companies and gas stations recognize that they have an opportunity to make higher profits by selling what gasoline they have at a higher price. These price increases will stimulate the quantity supplied and reduce the quantity demanded. As this occurs, the shortage will decrease. How far will the price rise? The price will rise until the shortage is eliminated and the quantity supplied equals quantity demanded. In other words, the market will be in equilibrium again. As before, the equilibrium occurs at a price of $1.40 per gallon and at a quantity of 600 gallons.

Generally any time the price for a good is below the equilibrium level, incentives built into the structure of demand and supply will create pressures for the price to rise. Similarly, any time the price for a good is above the equilibrium level, similar pressures will generally cause the price to fall.
As you can see, the quantity supplied or quantity demanded in a free market will correct over time to restore balance, or equilibrium.

**Equilibrium: Where Supply and Demand Intersect**

When two lines on a diagram cross, this intersection usually means something. On a graph, the point where the supply curve (S) and the demand curve (D) intersect is the **equilibrium.** The **equilibrium price** is the only price where the desires of consumers and the desires of producers agree—that is, where the amount of the product that consumers want to buy (quantity demanded) is equal to the amount producers want to sell (quantity supplied). This mutually desired amount is called the **equilibrium quantity.** At any other price, the quantity demanded does not equal the quantity supplied, so the market is not in equilibrium at that price. It should be clear from the previous discussions of surpluses and shortages, that if a market is not in equilibrium, market forces will push the market to the equilibrium.

If you have only the demand and supply schedules, and no graph, you can find the equilibrium by looking for the price level on the tables where the quantity demanded and the quantity supplied are equal (again, the numbers in bold in Table 1 indicate this point).

---

**Finding Equilibrium with Algebra**

We’ve just explained two ways of finding a market equilibrium: by looking at a table showing the quantity demanded and supplied at different prices, and by looking at a graph of demand and supply. We can also identify the equilibrium with a little algebra if we have equations for the supply and demand curves. Let’s practice solving a few equations that you will see later in the course. Right now, we are only going to focus on the math. Later you’ll learn why these models work the way they do, but let’s start by focusing on solving the equations. Suppose that the demand for soda is given by the following equation:

\[ Q_d = 16 - 2P \]

where \( Q_d \) is the amount of soda that consumers want to buy (i.e., quantity demanded), and \( P \) is the price of soda. Suppose the supply of soda is

\[ Q_s = 2 + 5P \]

where \( Q_s \) is the amount of soda that producers will supply (i.e., quantity supplied). (Remember, these are simple equations for lines). Finally, recall that the soda market converges to the point where supply equals demand, or

\[ Q_d = Q_s \]

We now have a system of three equations and three unknowns (\( Q_d, Q_s, \) and \( P \)), which we can solve with algebra. Since \( Q_d = Q_s \), we can set the demand and supply equations equal to each other:

\[ Q_d = Q_s \]

\[ 16 - 2P = 2 + 5P \]

**Step 1:** Isolate the variable by adding 2P to both sides of the equation, and subtracting 2 from both sides.

\[ 16 - 2P = 2 + 5P \]

\[ -2 + 2P = -2 + 2P \]

\[ 14 = 7P \]
Step 2: Simplify the equation by dividing both sides by 7.

\[
\begin{align*}
14 &= 7P \\
\frac{14}{7} &= \frac{7P}{7} \\
2 &= P
\end{align*}
\]

The equilibrium price of soda, that is, the price where \( Q_s = Q_d \) will be $2. Now we want to determine the quantity amount of soda. We can do this by plugging the equilibrium price into either the equation showing the demand for soda or the equation showing the supply of soda. Let’s use demand. Remember, the formula for quantity demanded is the following:

\[
Q_d = 16 - 2P
\]

Taking the price of $2, and plugging it into the demand equation, we get

\[
Q_d = 16 - 2(2) = 16 - 4 = 12
\]

So, if the price is $2 each, consumers will purchase 12. How much will producers supply, or what is the quantity supplied? Taking the price of $2, and plugging it into the equation for quantity supplied, we get the following:

\[
Q_s = 2 + 5P
\]

\[
Q_s = 2 + 5(2) = 2 + 10 = 12
\]

Now, if the price is $2 each, producers will supply 12 sodas. This means that we did our math correctly, since \( Q_d = Q_s \) and both \( Q_d \) and \( Q_s \) are equal to 12. That confirms that we’ve found the equilibrium quantity.

Watch It

Watch this video for a closer look at market equilibrium:
Watch this video online: https://youtu.be/W5nHpAn6FvQ

Equilibrium and Economic Efficiency

Equilibrium is important to create both a balanced market and an efficient market. If a market is at its equilibrium price and quantity, then it has no reason to move away from that point, because it’s balancing the quantity supplied and the quantity demanded. However, if a market is not at equilibrium, then economic pressures arise to move the market toward the equilibrium price and equilibrium quantity. This happens either because there is more supply than what the market is demanding or because there is more demand than the market is supplying. This balance is a natural function of a free-market economy.

Also, a competitive market that is operating at equilibrium is an efficient market. Economists typically define efficiency in this way: when it is impossible to improve the situation of one party without imposing a cost on another. Conversely, if a situation is inefficient, it becomes possible to benefit at least one party without imposing costs on others.
Efficiency:
equilibrium:
equilibrium price:
equilibrium quantity:
shortage (or excess demand):
surplus (or excess supply):

Figure 5. Demand and Supply for Gasoline: Equilibrium. At this equilibrium point, the market is efficient because the optimal amount of gasoline is being produced and consumed.

Efficiency in the demand and supply model has the same basic meaning: the economy is getting as much benefit as possible from its scarce resources, and all the possible gains from trade have been achieved. In other words, the optimal amount of each good and service is being produced and consumed. We will explore this important concept in detail in the next module on applications of supply and demand.

Try It
Visit this page in your course online to view this presentation.

Glossary

**efficiency:** when the optimal amount of goods are produced and consumed, minimizing waste

**equilibrium:** price and quantity combination where supply equals demand

**equilibrium price:** the (only) price where the quantity supplied in a market equals the quantity demanded

**equilibrium quantity:** the quantity both supplied and demanded at the equilibrium price

**shortage (or excess demand):** situation where the quantity demanded in a market is greater than the quantity supplied; occurs at prices above the equilibrium

**surplus (or excess supply):** situation where the quantity demanded in a market is less than the quantity supplied; occurs at prices below the equilibrium
CHANGES IN EQUILIBRIUM

Learning Objectives

- Create a graph that illustrates equilibrium price and quantity
- Predict how economic conditions cause a change in supply, demand, and equilibrium (using the four-step process)

Finding Equilibrium using the Four-Step Process

We know that equilibrium is the place where the supply and demand curves intersect, or the point where buyers want to buy the same amount that sellers want to sell. Let’s take a closer look at how to find the equilibrium point using the four-step process. These steps explain how to first, draw the demand a supply curves on a graph and find the equilibrium. Next, consider how an economic change (e.g. a natural disaster, a change in production technology, a change in tastes and preferences, income, etc.) might affect supply or demand, then make adjustments to the graph to identify the new equilibrium point.

**Step 1.** Draw demand and supply curves showing the market before the economic change took place. Think about the shift variables for demand, and the shift variables for supply. Using this diagram, find the initial equilibrium values for price and quantity.

**Step 2.** Decide whether the economic change being analyzed affects demand or supply. In other words, does the event refer to something in the list of demand shift variables or supply shift variables?

**Step 3.** Determine whether the effect on demand or supply causes the curve to shift to the right or to the left, and sketch the new demand or supply curve on the diagram. In other words, does the event increase or decrease the amount consumers want to buy or the amount producers want to sell?

**Step 4.** Identify the new equilibrium, and then compare the original equilibrium price and quantity to the new equilibrium price and quantity.

Let’s consider one example that involves a shift in supply and one that involves a shift in demand. Then we will consider an example where both supply and demand shift.

**Exercise 1: Good Weather for Salmon Fishing**

Let’s suppose that during the summer of 2015, weather conditions were excellent for commercial salmon fishing off the California coast. Heavy rains meant higher than normal levels of water in the rivers, which helps the salmon to breed. Slightly cooler ocean temperatures stimulated the growth of plankton, the microscopic organisms at the bottom of the ocean food chain, providing everything in the ocean with a hearty food supply. The ocean stayed calm during fishing season, so commercial fishing operations did not lose many days to bad weather. How did these climate conditions affect the quantity and price of salmon?

Let’s consider this situation using the four-step process and the data below.
Table 1. Salmon Fishing

<table>
<thead>
<tr>
<th>Price per Pound</th>
<th>Quantity Supplied in 2014</th>
<th>Quantity Supplied in 2015</th>
<th>Quantity Demanded</th>
</tr>
</thead>
<tbody>
<tr>
<td>$2.00</td>
<td>80</td>
<td>400</td>
<td>840</td>
</tr>
<tr>
<td>$2.25</td>
<td>120</td>
<td>480</td>
<td>680</td>
</tr>
<tr>
<td>$2.50</td>
<td>160</td>
<td>550</td>
<td>550</td>
</tr>
<tr>
<td>$2.75</td>
<td>200</td>
<td>600</td>
<td>450</td>
</tr>
<tr>
<td>$3.00</td>
<td>230</td>
<td>640</td>
<td>350</td>
</tr>
<tr>
<td>$3.25</td>
<td>250</td>
<td>670</td>
<td>250</td>
</tr>
<tr>
<td>$3.50</td>
<td>270</td>
<td>700</td>
<td>200</td>
</tr>
</tbody>
</table>

Let’s walk through the four steps together using this example, and see how the graph changes. Use the interactive graph below (Figure 1) by clicking on the arrows at the bottom of the activity to navigate through the steps. Visit this page in your course online to view this presentation.

**Figure 1 (Interactive Graph). Good Weather for Salmon Fishing: The Four-Step Process.**

In short, good weather conditions increased supply of the California commercial salmon. The result was a higher equilibrium quantity of salmon bought and sold in the market at a lower price.

Exercise 2: Newspapers and the Internet

According to the Pew Research Center for People and the Press, more and more people, especially younger people, are getting their news from online and digital sources. The majority of U.S. adults now own smartphones or tablets, and most of those Americans say they use them in part to get the news. From 2004 to 2012, the share of Americans who reported getting their news from digital sources increased from 24 percent to 39 percent. How has this trend affected consumption of print news media and radio and television news? Figure 2 and the text below illustrate the four-step analysis used to answer this question.

**Figure 2. Graph depicting the changing market for print news.**

**Step 1.** Draw a demand and supply model to think about what the market looked like before the event.
The demand curve $D_0$ and the supply curve $S_0$ show the original relationships. In this case, the curves are drawn without specific numbers on the price and quantity axis.
Step 2. Did the change described affect supply or demand?

Answer

A change in tastes, from traditional news sources (print, radio, and television) to digital sources, caused a change in demand for the former.

Step 3. Was the effect on demand positive or negative?

Answer

A shift to digital news sources will tend to mean a lower quantity demanded of traditional news sources at every given price, causing the demand curve for print and other traditional news sources to shift to the left, from D₀ to D₁.

Step 4. Compare the new equilibrium price and quantity to the original equilibrium price.

Answer

The new equilibrium (E₁) occurs at a lower quantity and a lower price than the original equilibrium (E₀).

The decline in print news reading predates 2004. Print newspaper circulation peaked in 1973 and has declined since then due to competition from television and radio news. In 1991, 55 percent of Americans indicated that they got their news from print sources, while only 29 percent did so in 2012. Radio news has followed a similar path in recent decades, with the share of Americans getting their news from radio declining from 54 percent in 1991 to 33 percent in 2012. Television news has held its own during the last fifteen years, with the market share staying in the mid to upper fifties. What does this suggest for the future, given that two-thirds of Americans under thirty years old say they don’t get their news from television at all?

Try It

Visit this page in your course online to view this presentation.

FINDING EQUILIBRIUM

Learning Objectives

- Explain what happens to supply, demand, and equilibrium when there is a change in both supply and demand

You have seen how changes in weather can influence supply and changes in consumer preferences can reduce demand, but what happens when both supply and demand are changing? Often changes in an economy affect both the
supply and the demand curves, making it more difficult to assess the impact on the equilibrium price. Let’s review one such example.

First, consider the following questions:

1. Suppose postal workers are successful in obtaining a pay raise from the U.S. Postal Service. Will this affect the supply or the demand for first-class mail? Why? Which determinant of demand or supply is being affected? Show graphically with before and after curves on the same axes. How will this change the equilibrium price and quantity of first-class mail?
2. How do you imagine the invention of email and text messaging affected the market for first-class mail? Why? Which determinant of demand or supply is being affected? Show graphically with before and after curves on the same axes. How will this change the equilibrium price and quantity of first-class mail?
3. Suppose that postal workers get a pay raise and email and text messaging become common. What will the combined impact be on the equilibrium price and quantity of first-class mail?

In order to complete a complex analysis like this it’s helpful to tackle the parts separately and then combine them, while thinking about possible interactions between the two parts that might affect the overall outcome. Let’s use the four-step process.

**Exercise: Postal Service**

**Part 1: A Pay Raise for Postal Workers**

**Step 1.** Draw a demand and supply model to illustrate what the market for the U.S. Postal Service looks like before this scenario starts. The demand curve D and the supply curve S show the original relationships.

**Step 2.** Will a pay raise for postal workers affect supply or demand?

**Answer**

**Figure 1. Money and Mail.** How do changes at the post office impact other aspects of the economy?

**Figure 2.** The supply and demand curves for the U.S. Postal Service after a pay raise for postal workers.
Labor compensation is a cost of production. A change in production costs cause a change in supply for the Postal Service.

**Step 3.** Is the effect on supply positive or negative?

**Answer**

Higher labor compensation leads to a lower quantity supplied of postal services at every given price, causing the supply curve for postal services to shift to the left, from $S$ to $S_1$.

**Step 4.** Compare the new equilibrium price and quantity to the original equilibrium price.

**Answer**

The new equilibrium occurs at a lower quantity and a higher price than the original equilibrium.

A pay raise for postal workers would represent an increase in the cost of production for the Postal Service. Production costs are a factor that influences supply; thus, the pay raise should decrease the supply of first-class mail, shifting the supply curve vertically by the amount of the pay raise. Intuitively, all else held constant, the Postal Service would like to charge a higher price that incorporates the higher cost of production. That is not to say the higher price will stick. From the graph (Figure 1), it should be clear that at that higher price, the quantity supplied is greater than the quantity demanded—thus there would be a surplus, indicating that the price the Postal Service desires is not an equilibrium price. Or to put it differently, at the original price ($P_1$), the decrease in supply causes a shortage driving up the price to a new equilibrium level ($P_2$). Note that the price doesn’t rise by the full amount of the pay increase. In short, a leftward shift in the supply curve causes a movement up the demand curve, resulting in a lower equilibrium quantity ($Q_2$) and a higher equilibrium price ($P_2$).

**Part 2: The Effect of Email and Text Messaging**

**Step 1.** We’ve already seen how a pay raise will shift the supply curve to the left. Now let’s consider how the invention of email and text messaging affects the market for first-class mail. Begin by drawing a demand and supply model reflecting this relationship.

---

**Figure 3.** The impact of email and text messages on the snail mail market.

**Step 2.** Does email and text messaging affect supply or demand?
A change in tastes away from snail mail toward digital messages will cause a change in demand for the Postal Service.

**Step 3.** Is the effect on demand positive or negative?

Answer

A change in tastes away from snail mail toward digital messages causes lower quantity demanded of postal services at every given price, causing the demand curve for postal services to shift to the left, from D to D1.

**Step 4.** Compare the new equilibrium price and quantity to the original equilibrium price.

Answer

The new equilibrium occurs at a lower quantity and a lower price than the original equilibrium.

To summarize, since many people find email and texting more convenient than sending a letter, we can assume that tastes and preferences for first-class mail will decline. This decrease in demand is shown by a leftward shift in the demand curve and a movement along the supply curve, which creates a surplus in first-class mail at the original price (shown as $P_2$). The shortage causes a decrease in the equilibrium price (to $P_3$) and a decrease in the equilibrium quantity (to $Q_3$). Intuitively, less demand for first-class mail leads to a lower equilibrium quantity and (ceteris paribus) a lower equilibrium price.

**Part 3: Combining Factors**

Parts 1 and 2 are straightforward, but when we put them together it becomes more complex. Think about it this way: in Part 1, the equilibrium quantity fell due to decreased supply. In Part 2, the equilibrium quantity also fell, this time due to the decreased demand. So putting the two parts together, we would expect to see the final equilibrium quantity ($Q_3$) to be smaller than the original equilibrium quantity ($Q_1$). So far, so good.

Now consider what happens to the price. In Part 1, the equilibrium price increased due to the reduction in supply. But in Part 2, the equilibrium price decreased due to the decrease in demand! What will happen to the equilibrium price? The net effect on price can’t be determined without knowing which curve shifts more, demand or supply. The equilibrium price could increase, decrease, or stay the same. You just can’t tell from graphical analysis alone. This is not unusual. When both curves shift, typically we can determine the overall effect on price or on quantity, but not on both. In this case, we determined the overall effect on the equilibrium quantity, but not on the equilibrium price.
Figure 4. Decreased supply and decreased demand cause the equilibrium quantity to fall ($Q_3$ is be smaller than the original equilibrium quantity of $Q_1$). It is hard to pinpoint what happens to the price, however, as it may decline with the shrinking demand, but also increase with production costs.

Try It
Visit this page in your course online to view this presentation.

CHANGES IN SUPPLY AND DEMAND

Learning Objectives

- Describe the differences between changes in demand and changes in the quantity demanded
- Describe the differences between changes in supply and changes in quantity supplied

It’s hard to overstate the importance of understanding the difference between shifts in curves and movements along curves. Remember, when we talk about changes in demand or supply, we do not mean the same thing as changes
in quantity demanded or quantity supplied.

A change in demand refers to a shift in the entire demand curve, which is caused by a variety of factors (preferences, income, prices of substitutes and complements, expectations, population, etc.). In this case, the entire demand curve moves left or right:

![Figure 1. Change in Demand](image1)

**Figure 1. Change in Demand.** A change in demand means that the entire demand curve shifts either left or right. The initial demand curve $D_0$ shifts to become either $D_1$ or $D_2$. This could be caused by a shift in tastes, changes in population, changes in income, prices of substitute or complement goods, or changes future expectations.

A change in quantity demanded refers to a movement along the demand curve, which is caused only by a chance in price. In this case, the demand curve doesn’t move; rather, we move along the existing demand curve:

![Figure 2. Change in Quantity Demanded](image2)

**Figure 2. Change in Quantity Demanded.** A change in the quantity demanded refers to movement along the existing demand curve, $D_0$. This is a change in price, which is caused by a shift in the supply curve.

Similarly, a change in supply refers to a shift in the entire supply curve, which is caused by shifter such as taxes, production costs, and technology. Just like with demand, this means that the entire supply curve moves left or right:
A change in supply means that the entire supply curve shifts either left or right. The initial supply curve $S_0$ shifts to become either $S_1$ or $S_2$. This is caused by production conditions, changes in input prices, advances in technology, or changes in taxes or regulations.

A change in quantity supplied refers to a movement along the supply curve, which is caused only by a change in price. Similar to demand, a change in quantity supplied means that we’re moving along the existing supply curve:

Here’s one way to remember: a movement along a demand curve, resulting in a change in quantity demanded, is always caused by a shift in the supply curve. Similarly, a movement along a supply curve, resulting in a change in quantity supplied, is always caused by a shift in the demand curve.

Watch It

Watch this video for another demonstration of the important distinction between these terms.
Watch this video online: https://youtu.be/BwNzl15NOTI
Try It

Try graphing each of these situations to determine if they cause a shift in demand, quantity demanded, supply, or quantity supplied.

Visit this page in your course online to view this presentation.

Visit this page in your course online to view this presentation.

Visit this page in your course online to view this presentation.

Visit this page in your course online to view this presentation.

Glossary

demand: the relationship between the price and the quantity demanded of a certain good or service

quantity demanded: the total number of units of a good or service consumers are willing to purchase at a given price

quantity supplied: the total number of units of a good or service producers are willing to sell at a given price

shift in demand: when a change in some economic factor (other than price) causes a different quantity to be demanded at every price

shift in supply: when a change in some economic factor (other than price) causes a different quantity to be supplied at every price

supply: the relationship between price and the quantity supplied of a certain good or service

LEARN BY DOING: FOOD TRUCKS AND CHANGES IN EQUILIBRIUM

Try It

In the following simulation, you will have the opportunity to change the weather and/or change the cost of parking in order to push up the price of food from a food truck. You can play the simulation multiple times to see how different
choices lead to different outcomes.
Visit this page in your course online to view this presentation.

LEARN BY DOING: CALCULATING EQUILIBRIUM

Try It

These questions allow you to get as much practice as you need, as you can click the link at the top of the first question ("Try another version of these questions") to get a new set of questions. Practice until you feel comfortable doing the questions and then move on.
Note that you will use the information provided in the first question for all of the questions on this page.
Visit this page in your course online to practice before taking the quiz.

PUTTING IT TOGETHER: SUPPLY AND DEMAND

The demand and supply model emphasizes that prices are not set only by demand or only by supply, but by the interaction between the two. In 1890, the famous economist Alfred Marshall wrote that asking whether supply or demand determined a price was like arguing "whether it is the upper or the under blade of a pair of scissors that cuts a piece of paper." The answer is that both blades of the demand and supply scissors are always involved.

You'll remember that we started this module by considering changes in global coffee prices. Let's focus specifically on the drought of 2014 and see how the drought in Brazil affected supply and demand for coffee.

Now that we understand more about supply and demand, we can answer a few important questions: How does a drought impact supply? What impact will the quantity supplied have on the equilibrium price?
In 2014, the coffee regions of Brazil experienced a serious drought. The lack of rain in Brazil's coffee-growing region delayed the tree-flowering period, which spans October and November. When the trees don't flower, they don't produce coffee. Weather conditions also affect the pollination of coffee trees that have already flowered: drought makes the blooms very delicate, which can cause them to fall off the tree. In 2014, the combined impact of these consequences meant a 13 percent drop in production from the previous year, to only 48 million 60-kilogram bags. (Note: http://www.wallstreetdaily.com/2014/10/21/coffee-prices-brazil-drought/)

These are poor natural conditions for coffee growers, and they cause a reduction in the supply. Graphically, such a reduction means a shift to the left in the supply curve (shown in Figure 3, below), indicating that suppliers are providing less coffee at every price.
Figure 3. Poor weather conditions results in a leftward shift in the supply curve.

We can see that this shift in the supply curve will change the quantity supplied and the equilibrium price. At the original price ($P_1$), the decrease in supply causes a shortage—more people want coffee at that low price than the suppliers are able to provide. This drives up the price to a new equilibrium level ($P_2$). In short, a leftward shift in the supply curve causes a movement up the demand curve, resulting in a lower equilibrium quantity ($Q_2$) and a higher equilibrium price ($P_2$).

This impact is clear in an economic model like the graph above, but does it really affect consumers? Absolutely!—during this period, Starbucks raised its prices by 8 percent, and Folgers raised its prices by 9 percent. Coffee retailers were able to limit some of the impact of the rising coffee prices by drawing down their stock of green beans that were purchased before the drought and passing on some of the cost on to their customers as a higher price. (Note: http://www.barrons.com/articles/rise-in-coffee-prices-nearing-peak-1444457073) Some people—call them the coffee addicts—continue to drink coffee and pay the higher price. Others switch to tea or soft drinks. No government commission is needed to figure out how to adjust coffee prices, which companies will be allowed to process the remaining supply, which supermarkets in which cities will get how much coffee to sell, or which consumers will ultimately be allowed to drink the brew. Such adjustments in response to price changes happen all the time in a market economy, often so smoothly and rapidly that we barely notice them.

Think for a moment of all the seasonal foods that are available and inexpensive at certain times of the year, like fresh corn in midsummer, but more expensive at other times of the year. People alter their diets and restaurants alter their menus in response to these fluctuations in prices without fuss or fanfare. For both the U.S. economy and the world economy as a whole, markets—that is, demand and supply—are the primary social mechanism for answering the basic questions about what is produced, how it is produced, and for whom it is produced.

It's very common to see the impact of drought and other natural factors on supply, equilibrium quantity, and equilibrium price. The following video provides a brief example in the United States.

Watch It

Watch this video online: https://youtu.be/xBU8Z0Accds
Why evaluate the applications of supply and demand?

In the module on supply and demand, we defined a free market as one with no government intervention. In this module, we will learn about the applications of supply and demand to explore the outcomes, both anticipated and otherwise, when government does intervene in a market.

Economists believe there are a small number of fundamental principles that explain how economic agents respond in different situations. Two of these principles, which we have already been introduced to, are the laws of demand and supply.

Governments can pass laws affecting market outcomes, but no law can negate these economic principles. Rather, the principles will manifest themselves in sometimes unanticipated ways, which may subvert the intent of the government policy. This is one of the major conclusions of this module.

The three best examples of this are:

- Price floors—a legal minimum price in a market, e.g. the minimum wage;
- Price ceilings—a legal maximum price in a market, e.g. rent controls in certain cities;
- Tax incidence—who ends up paying a tax? For example, if the local government adds a sales tax on restaurant meals, is the tax paid by the diners or does it come out of the restaurant’s profits? (We’ll study more about this in the next module on elasticity.)

Understanding all the effects, both anticipated and unanticipated, of government intervention in a market is critical to determining whether the policy achieves its goal.

As you go through this module, make sure to keep in mind who is a given policy (e.g., a minimum wage) supposed to help? Only then can you evaluate whether the policy is a good one or not. Check out the following video about the minimum wage.

Watch It

The story dates from when the minimum wage was $5.15, but the issues are still relevant today as the discussion continues about further increases in the minimum wage.

Watch this video online: https://youtu.be/ZypGjooHxY

Click through this exercise to briefly learn about two ways that government frequently interferes with a market economy and the consequences these actions have on the market.
INTRODUCTION TO PRICE CEILINGS AND PRICE FLOORS

What you’ll learn to do: analyze the economic effect of government setting price ceilings and floors

In this section, we will explore the outcomes, both anticipated and otherwise, when government intervenes in a market either to prevent the price of some good or service from rising “too high” or to prevent the price of some good or service from falling “too low.”

First, we will take a look at what happens when prices are held below the equilibrium level. Governments typically set a price ceiling to protect consumers by making necessary products affordable, but you’ll come to see how this sometimes backfires by creating a market shortage.

Next, we will see what happens when a price floor forces prices above a minimum standard, such as a minimum wage. While a minimum wage seems like a great benefit for workers, you’ll see that some effects of a minimum wage can
actually hurt those in the workforce.

PRICE CEILINGS

Learning Objectives

- Analyze the consequences of the government setting a binding price ceiling, including the economic impact on price, quantity demanded and quantity supplied
- Compute and demonstrate the market shortage resulting from a price ceiling

Supply and Demand Model

Economists believe there are a small number of fundamental principles that explain how economic agents respond in different situations. Two of these principles, which we have already introduced, are the laws of supply and demand. Governments can pass laws affecting market outcomes, but no law can negate these economic principles. Rather, the laws of supply and demand often become apparent in sometimes unexpected ways, which may undermine the intent of the government policy. This is one of the major conclusions of this section.

Controversy sometimes surrounds the prices and quantities established by supply and demand, especially for products that are considered necessities. In some cases, discontent over prices turns into public pressure on politicians, who may then pass legislation to prevent a certain price from climbing “too high” or falling “too low.”

Watch It

Watch this video to see a historical example of what happened to the U.S. economy because of government-enacted price controls in the 1970s.
Watch this video online: https://youtu.be/sq1zlj8s8R0

The supply and demand model shows how people and firms will react to the incentives that laws provide to control prices, in ways that will often lead to undesirable consequences. Alternative policy tools can often achieve the desired goals of price control laws, while avoiding at least some of their same costs and tradeoffs.

Price Ceilings

Laws that government enacts to regulate prices are called Price Controls. Price controls come in two flavors. A price ceiling keeps a price from rising above a certain level (the “ceiling”), while a price floor keeps a price from falling below a certain level (the “floor”). First, let’s use the supply and demand framework to analyze price ceilings.

A price ceiling is a legal maximum price that one pays for some good or service. A government imposes price ceilings in order to keep the price of some necessary good or service affordable. For example, in 2005 during Hurricane Katrina,
the price of bottled water increased above $5 per gallon. As a result, many people called for price controls on bottled water to prevent the price from rising so high. In this particular case, the government did not impose a price ceiling, but there are other examples of where price ceilings did occur.

In many markets for goods and services, demanders outnumber suppliers. Consumers, who are also potential voters, sometimes unite behind a political proposal to hold down a certain price. In some cities, such as Albany, renters have pressed political leaders to pass rent control laws, a price ceiling that usually works by stating that rents can be raised by only a certain maximum percentage each year. Some of the best examples of rent controls occur in urban areas, such as New York, Washington D.C., or San Francisco.

Rent control becomes a politically hot topic when rents begin to rise rapidly. Everyone needs an affordable place to live. Perhaps a change in tastes makes a certain suburb or town a more popular place to live. Perhaps locally-based businesses expand, bringing higher incomes and more people into the area. Changes of this sort can cause a change in the demand for rental housing. The interactive graph below (Figure 1) explains how this happens.

Visit this page in your course online to view this presentation.

**Figure 1 (Interactive Graph). A Price Ceiling Example—Rent Control.**

The following table shows the changes in quantity supplied and quantity demanded at each price for the above graphs.

<table>
<thead>
<tr>
<th>Price</th>
<th>Original Quantity Supplied</th>
<th>Original Quantity Demanded</th>
<th>New Quantity Demanded</th>
</tr>
</thead>
<tbody>
<tr>
<td>$400</td>
<td>12,000</td>
<td>18,000</td>
<td>23,000</td>
</tr>
<tr>
<td>$500</td>
<td>15,000</td>
<td>15,000</td>
<td>19,000</td>
</tr>
<tr>
<td>$600</td>
<td>17,000</td>
<td>13,000</td>
<td>17,000</td>
</tr>
<tr>
<td>$700</td>
<td>19,000</td>
<td>11,000</td>
<td>15,000</td>
</tr>
<tr>
<td>$800</td>
<td>20,000</td>
<td>10,000</td>
<td>14,000</td>
</tr>
</tbody>
</table>

In the graphs above, we saw what happens when a rent control law is passed to keep the price at the original equilibrium of $500 for a typical apartment. The horizontal line at the price of $500 shows the legally fixed maximum price set by the rent control law. However, the underlying forces that shifted the demand curve to the right are still there. At that price ($500), the quantity supplied remains at the same 15,000 rental units, but the quantity demanded is 19,000 rental units. In other words, the quantity demanded exceeds the quantity supplied, so there is a shortage of rental housing. One of the ironies of price ceilings is that while the price ceiling was intended to help renters, there are actually fewer apartments rented out under the price ceiling (15,000 rental units) than would be the case at the market rent of $600 (17,000 rental units). When a price ceiling is set below the equilibrium price, as in this example, it is considered a **binding price ceiling**, thereby resulting in a shortage.

Price ceilings do not simply benefit renters at the expense of landlords. Rather, some renters (or potential renters) lose their housing as landlords convert apartments to co-ops and condos. Even when the housing remains in the rental market, landlords tend to spend less on maintenance and on essentials like heating, cooling, hot water, and lighting. The first rule of economics is you do not get something for nothing—everything has an opportunity cost. So if renters get “cheaper” housing than the market requires, they tend to also end up with lower quality housing.

Price ceilings have been proposed for other products, for example, for prescription drugs, doctor and hospital fees, the charges made by some automatic teller bank machines, and auto insurance rates. The general results of any price ceiling are the same: price ceilings are enacted in an attempt to keep prices low for those who need the product. But when the market price is not allowed to rise to the equilibrium level, quantity demanded exceeds quantity supplied, and thus a shortage occurs. Those who manage to purchase the product at the lower price given by the price ceiling will benefit, but sellers of the product will suffer, along with those who are not able to purchase the product at all. To the extent that producers cannot easily reduce the quantity supplied, they will tend to allow the quality to decline.
Watch It

The following video explores the effects of price ceilings. The speakers identify five major consequences:

1. Shortages
2. Reduced quality
3. Wasted time and resources
4. Deadweight loss, or a loss of gains from trade
5. Misallocation of resources

The first two consequences are explained in the video. We’ll address the others later in the module in the discussion of efficiency.

Watch this video online: https://youtu.be/RBGHmCIBr9M

Glossary

binding price ceiling when a price ceiling is set below the equilibrium price, resulting in a shortage
price ceiling: a legal maximum price for a product
price floor: a legal minimum price for a product

Learning Objectives

- Analyze the consequences of the government setting a binding price floor, including the economic impact on price, quantity demanded and quantity supplied
- Compute and demonstrate the market surplus resulting from a price floor
A **price floor** is the lowest price that one can legally charge for some good or service. Perhaps the best-known example of a price floor is the minimum wage, which is based on the view that someone working full time should be able to afford a basic standard of living. The federal minimum wage in 2016 was $7.25 per hour, although some states and localities have a higher minimum wage. The federal minimum wage yields an annual income for a single person of $15,080, which is slightly higher than the Federal poverty line of $11,880. As the cost of living rises over time, the Congress periodically raises the federal minimum wage.

Price floors are sometimes called “price supports,” because they support a price by preventing it from falling below a certain level. Around the world, many countries have passed laws to create agricultural price supports. Farm prices and thus farm incomes fluctuate, sometimes widely. Even if, on average, farm incomes are adequate, some years they can be quite low. The purpose of price supports is to prevent these swings.

The most common way price supports work is that the government enters the market and buys up the product, adding to demand to keep prices higher than they otherwise would be. According to the Common Agricultural Policy reform passed in 2013, the European Union (EU) will spend about 60 billion euros per year, or 67 billion dollars per year (with the November 2016 exchange rate), or roughly 38% of the EU budget, on price supports for Europe’s farmers from 2014 to 2020.

Figure 2 illustrates the effects of a government program that assures a price above the equilibrium by focusing on the market for wheat in Europe. In the absence of government intervention, the price would adjust so that the quantity supplied would equal the quantity demanded at the equilibrium point $E_0$, with price $P_0$ and quantity $Q_0$. However, policies to keep prices high for farmers keeps the price above what would have been the market equilibrium level—the price $P_f$ shown by the dashed horizontal line in the diagram. The result is a quantity supplied in excess of the quantity demanded ($Q_d$). When quantity supplied exceeds quantity demanded, a surplus exists. When a price floor is set above the equilibrium price, as in this example, it is considered a **binding price floor**.

![Figure 1. Protesters call for a $15 minimum wage.](image1)

Economists estimate that the high-income areas of the world, including the United States, Europe, and Japan, spend roughly $1 billion per day in supporting their farmers. If the government is willing to purchase the excess supply (or to provide payments for others to purchase it), then farmers will benefit from the price floor, but taxpayers and consumers of food will pay the costs through higher taxes and higher prices. Agricultural economists and policy makers have
offered numerous proposals for reducing farm subsidies. In many countries, however, political support for subsidies for farmers remains strong. This is either because the population views this as supporting the traditional rural way of life or because of industry’s lobbying power of the agro-business.

**Try It**

Visit this page in your course online to view this presentation.

**Try It**

This next question allow you to get as much practice as you need, as you can click the link at the top of the question (“Try another version of this question”) to get a new version of the question. Practice until you feel comfortable with this concept.

Visit this page in your course online to practice before taking the quiz.

**Watch It**

We mentioned earlier that the minimum wage is a good example of a price floor, since employers are required to pay no less than the minimum wage for workers. The following video makes a strong case for why a minimum wage causes a surplus of labor, i.e. unemployment. As you watch it, consider what you think about a minimum wage.

Watch this video online: [https://youtu.be/IFbYM2EDz40](https://youtu.be/IFbYM2EDz40)

**Do price ceilings and floors change demand or supply?**

Neither price ceilings nor price floors cause demand or supply to change. They simply set a price that limits what can be legally charged in the market. Remember, changes in price do not cause demand or supply to change. In other words, they do not change the equilibrium. Price ceilings and price floors can cause a different choice of quantity demanded along a demand curve, but they do not move the demand curve. Price controls can cause a different choice of quantity supplied along a supply curve, but they do not shift the supply curve.

**Glossary**

- **binding price floor** when a price floor is set above the equilibrium price and results in a surplus
- **price ceiling**: a legal maximum price
- **price control**: government laws to regulate prices instead of letting market forces determine prices
- **price floor**: a legal minimum price for a product
LEARN BY DOING: SHORTAGE AND SURPLUS

Try It

Rather than have you read more about shortage and surplus, we’d prefer to have you practice what you’ve learned so far and see for yourself if you understand it.

These questions allow you to get as much practice as you need, as you can click the link at the top of the first question (“Try another version of these questions”) to get a new version of the questions. Practice until you feel comfortable doing these questions.

Note that you’ll use the information provided in the first question for all of the questions on this page.

Visit this page in your course online to practice before taking the quiz.

A CLOSER LOOK AT PRICE CONTROLS

Learning Objectives

- Analyze the economic effect of government setting price ceilings and floors

Price Controls

People often expect the government to solve problems that they seem unable to solve on their own. Sometimes this is effective and sometimes it is not. Price controls, either price ceilings or price floors, often have unanticipated side effects. Think about it—passing a law doesn’t by itself make economic problems go away!

Such is the case with claims of price gouging or the charging of “excessively high” prices, as is sometimes reported in the wake of natural disasters, like the need for water after Hurricane Katrina or the need for gas following Hurricane Harvey. Imposing a price ceiling below the equilibrium price may create as many problems as it solves. The problem originates from the fact that the demand for the good increases suddenly and dramatically. After Katrina, freshwater supplies were compromised and bottled water was hard to access because of the storm damage, so the price of bottled water increased above $5 per gallon. while the supply of bottled water was less as a result of storm damage. The question is how to deal with the shortage, that is, how to allocate the limited supply of bottled water among competing needs and wants. Figure 1 depicts the example of a hypothetical price control for water after Hurricane Katrina.
Figure 1. Price control for water after Hurricane Katrina. D and S show the original demand and supply curves, with equilibrium at $1. After the hurricane, demand for water shifted to the right at D’. Similarly, the limited supply of water shifted the supply curve to the left at S’. While the new equilibrium price would be $7, the price ceiling of $2 results in a shortage.

When a price ceiling reduces the legal price of a product, businesses have less incentive to supply the product. Economically speaking, the law of supply says that at lower prices, the quantity supplied will be lower. At the same time, the law of demand states that at a lower price, the quantity demanded will be higher. This can be seen clearly in the graph. Suppose there is a price ceiling at $2 per gallon of bottled water. The quantity demanded is shown in the figure as Qd. The quantity supplied is shown as Qs. Since Qs < Qd, there is a shortage. So who gets the limited supply?

Unfortunately, there is no clear answer to this. It could be first come, first serve. It could be friends of the seller. In many cases, what results are under-the-table payments by consumers willing to violate the law. What is certain is that less bottled water gets to consumers than would be the case if the price were allowed to rise. Many would argue that this shortfall is not the best outcome.

Link It Up

Click here to watch a relatively recent example of food shortages in Venezuela.

Try It

Visit this page in your course online to view this presentation.

Try It

This next question allow you to get as much practice as you need, as you can click the link at the top of the question (“Try another version of this question”) to get a new version of the question. Practice until you feel comfortable with this concept.

Visit this page in your course online to practice before taking the quiz.

Licensing & Attributions

CC licensed content, Original

INTRODUCTION TO SURPLUS

What you’ll learn to do: define, calculate, and illustrate consumer, producer, and total surplus

Earlier in this course we introduced the concept of efficiency and pointed out that there are several types. Productive efficiency means producing the most output possible with the available resources. In other words, it means producing without waste. If you recall the production possibilities frontier, operating inside the frontier means the society is not producing efficiently, since all resources are not being used. Productive efficiency occurs only on the PPF.

Figure 1. Productive and Allocative Efficiency. This graph shows the production possibilities frontier for education and healthcare. All choices along the PPF (points A, B, C, D, and F) display productive efficiency. Any point inside the production possibilities frontier (R) is productively inefficient and wasteful because it’s possible to produce more of one good, the other good, or some combination of both goods.
But there are an infinite number of points on the PPF. What is the optimal point on the PPF, or what is the optimal quantity of each good for society to produce? The answer to this critically important question is given by allocative efficiency. Allocative efficiency maximizes the net social benefit of some product. These same ideas about efficiency can be applied to individual markets. When markets are free and competitive, equilibrium results in the efficient amount of a good or service is produced. By contrast, anytime there is a price ceiling or price floor, or when market participants do not buy and sell at the equilibrium price, the amount of the product being supplied will be inefficient, and society will suffer a deadweight loss.

TRADE AND EFFICIENCY

Learning Objectives

- Explain why voluntary trade benefits both parties and why it leads to allocative efficiency

Getting a Good Deal or Making a Good Deal

Why do people make transactions? Is it because the seller has a surplus of goods or the buyer has a shortage of them? Not exactly. The short answer is that people make transactions because they value the same goods differently at the margin. Remember that marginal analysis involves weighing the benefits and costs of choosing a little bit more or a little bit less of a good.

Suppose Bill loves to snack on apples, while Angie thinks apples are just okay. Suppose they each have a basket containing a dozen apples. Because Bill loves apples, he places a higher value on one more apple than Angie does. That's what “at the margin” means. Bill is considering one more apple. Suppose Bill thinks another apple would be worth $1.00, while Angie thinks another apple is only worth $0.10. If Bill offered to buy an apple for $0.50 from Angie, would she agree to the transaction?

Since Angie thinks the apple is only worth $0.10, then it would be to her advantage to sell one to Bill and use the $0.40 profit for something she values more than apples. Would Bill benefit from the deal? Since he thinks an apple is worth a dollar, if he could get it for fifty cents, he would be making $0.50 profit. If two parties differ on what some good is worth, they can each benefit from trading the good from the person who values it less to the person who values it more.

If trading one apple is good for both parties, would trading more be better? What motivated the transaction in the first place? It was the difference in opinion between Bill and Angie about what an apple is worth. The value one places on an item depends on tastes in general (in this case it was taste for apples), and how much more of a good a person would like (or how many apples were already consumed). If Angie is very hungry, it’s likely she would value an apple more
than normal. Similarly, if Bill had just eaten five apples, he probably would value one more less than he normally values apples.

This suggests another idea we've looked at before: the **law of diminishing marginal utility**. Because of diminishing marginal returns, the more of something you already have, the less one more unit is worth to you. Thus, we can graph Bill's marginal value curve as shown in Figure 1. Similarly, Angie's marginal value curve has a similar shape, but it's lower on the graph to reflect the fact that Angie likes apples less than Bill does.

![Figure 1](image)

*Figure 1. Bill and Angie each have a basket with 12 apples. Bill is at B₁ and Angie is at A₁. Bill likes apples more than Angie. For Bill, the 12th apple is worth $1.00, while for Angie, it's worth only $0.10.*

**Try It**

Visit this page in your course online to view this presentation.

**Trade and Efficiency**

What this means is that the more apples Bill has, the less he values another. Similarly, the less apples Angie has, the more she values one more. Thus, as Angie sells more apples to Bill, her marginal value increases while his decreases. That suggests an answer to the question posed above: Bill and Angie should keep trading apples until they place the same value on them. This is shown in Figure 2, where Bill has bought three apples from Angie. At that point, they will have maximized the benefits from trading apples. Economists describe these benefits from trading as an improvement in **allocative efficiency**. (Note: This page summarizes ideas from Chapter 3 of Armen A. Alchian & William R. Allen, *Exchange & Production: Competition, Coordination, & Control*, Wadsworth Publishing Company, Belmont, California. 1983.)
**allocative efficiency**: when benefits of trade are maximized and the mix of goods being produced represents the mix that society most desires.

**law of diminishing marginal utility**: as we consume more of a good or service, the utility we get from additional units of the good or service tend to become smaller than what we received from earlier units.

**marginal analysis**: comparing the benefits and costs of choosing a little more or a little less of a good.

---

**Try It**

Visit this page in your course online to view this presentation.

---

**Glossary**

---

**CONSUMER & PRODUCER SURPLUS**
Learning Objectives

- Explain, calculate, and illustrate consumer surplus
- Explain, calculate, and illustrate producer surplus
- Explain, calculate, and illustrate social surplus

Demand, Supply and Efficiency

The familiar demand and supply diagram holds within it the concept of allocative efficiency. One typical way that economists define efficiency is when it is impossible to improve the situation of one party without imposing a cost on another. Conversely, if a situation is inefficient, it becomes possible to benefit at least one party without imposing costs on others.

Efficiency in the demand and supply model has the same basic meaning: the economy is getting as much benefit as possible from its scarce resources and all the possible gains from trade have been achieved. In other words, the optimal amount of each good and service is being produced and consumed.

Consumer Surplus, Producer Surplus, Social Surplus

Consider a market for tablet computers, as shown in Figure 1. We usually think of demand curves as showing what quantity of some product consumers will buy at any price, but a demand curve can also be read the other way. If we choose a quantity of output, the demand curve shows the maximum price consumers would be willing to pay for that quantity. According to the demand curve in Figure 1, if producers wanted to sell a quantity of 20 million tablets, some customers are willing to pay $90 each (see point J.) In other words, a tablet is worth $90 to those customers.

However, that doesn’t mean that those customers will end up paying $90. Figure 1 shows that the equilibrium price is $80 and the equilibrium quantity is 28 million tablets. At that price, each customer who would have been willing to pay $90 for a tablet is getting a good deal. We all know what a good deal is—it’s when you get something for less than you think it’s worth. We don’t have to stop there. If suppliers chose to produce only 14 tables (as shown in point K), we can...
look at Figure 1 and up to the demand curve to see that some customers would have been willing to pay about $115 for a tablet at this quantity produced. What that means is that this subset of customers got an even better deal at the equilibrium price.

The demand curve shows what consumers are willing to pay for any given quantity of tablets. In other words, the height of the demand curve at any quantity shows what some consumers think those tablets are worth. We can formalize this idea of how good a deal consumers get on a transaction using the concept of **consumer surplus**.

Since a demand curve traces consumers' willingness to pay for different quantities, we can define the gain to consumers as the difference between what they would have been willing to pay and the price that they actually paid. At point J, consumers were willing to pay $90, but they were able to purchase tablets at the equilibrium price of $80, so they gained $10 of extra value on each tablet. This is exactly analogous to the “profit” Bill earned from buying apples that we described in the previous page of reading. If we add up the gains at every quantity, we can measure the consumer surplus as the area under the demand curve up to the equilibrium quantity and above the equilibrium price. In Figure 1, the consumer surplus is the area labeled F.

The supply curve shows the quantity that firms are willing to supply at each price. For example, point K in Figure 1 illustrates that firms would have been willing to supply a quantity of 14 million tablets at a price of $45 each. Those producers were instead able to charge the equilibrium price of $80, clearly receiving an extra benefit beyond what they required to supply the product. The amount that a seller is paid for a good minus the seller's actual cost is called **producer surplus**. In Figure 1, producer surplus is the area labeled G—that is, the area between the market price and the segment of the supply curve below the equilibrium.

To summarize, producers created and sold 28 tablets to consumers. Both producers and consumers benefited. The value of the tablets is the area under the demand curve up to the equilibrium quantity. The cost to produce that value is the area under the supply curve. The new value created by the transactions, i.e. the net gain to society, is the area between the supply curve and the demand curve, that is, the sum of producer surplus and consumer surplus. This sum is called **social surplus**, also referred to as economic surplus or total surplus. In Figure 1 we show social surplus as the area F + G. Social surplus is larger at the equilibrium quantity and price than it would be at any other quantity. This is what economists mean when they say that market equilibrium is (perfectly) allocatively efficient. At the efficient level of output, it is impossible to produce greater consumer surplus without reducing producer surplus, and it is impossible to produce greater producer surplus without reducing consumer surplus. In other words, the consumer and producers gains from exchange are maximized at the equilibrium point.

### Try It

Visit this page in your course online to view this presentation.

### Watch It

In this video, you'll consider the holiday market for Santa hats. The market is efficient and both consumer and producer surplus are maximized at the equilibrium point of $5.

If the government establishes a price ceiling, a shortage results, which also causes the producer surplus to shrink, and results in inefficiency called **deadweight loss**.

If government implements a price floor, there is a surplus in the market, the consumer surplus shrinks, and inefficiency produces deadweight loss.

Watch this video online: [https://youtu.be/n0LXkA9kato](https://youtu.be/n0LXkA9kato)

### Example: Calculate consumer surplus
Figure 2. Consumer and producer surpluses are shown as the area where consumers would have been willing to pay a higher price for a good or the price where producers would have been willing to sell a good.

In the sample market shown in the graph, equilibrium price is $10 and equilibrium quantity is 3 units. The consumer surplus area is highlighted above the equilibrium price line. This area can be calculated as the area of a triangle. Recall that to find the area of a triangle, you will need to know its base and height. Refer to the following example if you need a refresher.

Let's apply the calculation for the area of a triangle to our example market to see the added value that consumers will get for this item at the equilibrium price in our sample market.

**Step 1:** Define the base and height of the consumer surplus triangle.

\[
\text{Area} = \frac{1}{2} \text{base} \times \text{height}
\]

\[
= \frac{1}{2} (4) \times (3)
\]

\[
= \frac{1}{2} (12)
\]

\[
= 6
\]

Figure 3. The area of a triangle.
The base of the consumer surplus triangle is 3 units long. Be careful when you define the height of this triangle, it is tempting to say it is 25, can you see why it isn’t? The height is determined by the distance from the equilibrium price line and where the demand curve intersects the vertical axis. The height of the triangle begins at $10 and ends at $25, so it will be $25 – $10 = $15

\[ b = 3 \]
\[ h = 15 \]

**Step 2:** Apply the values for base and height to the formula for the area of a triangle.

\[ A = \frac{1}{2} b \times h \]
\[ A = \frac{1}{2} \times 15 \]
\[ A = \frac{1}{2} \times 45 \]
\[ A = \frac{45}{2} = 22.5 \]

---

**Glossary**

deadweight loss: the loss in social surplus that occurs when a market produces an inefficient quantity
producer surplus: the value to producers of their sales above their cost of production
social (or economic or total) surplus: the sum of consumer and producer surplus at some quantity and price of output

---

**Licensing & Attributions**

CC licensed content, Original
- Modification, adaptation, and original content. **Provided by:** Lumen Learning. **License:** CC BY: Attribution

CC licensed content, Shared previously
- Principles of Demand, Supply, and Efficiency. **Authored by:** OpenStax College. **Located at:** https://cnx.org/contents/vEmOH_-p@4.44:yi4Yqqa@2/Demand-Supply-and-Efficiency. **License:** CC BY: Attribution. **License Terms:** Download for free at http://cnx.org/contents/bc498e1f-efe9-43a0-8dea-d3569ad09a82@4.44

All rights reserved content
- Deadweight Loss. **Authored by:** ACDCLeadership. **Located at:** https://www.youtube.com/watch?v=mOLkkA9kato&list=PL6B2DBE4C2FC8F845&index=12. **License:** Other. **License Terms:** Standard YouTube License

---

**INEFFICIENCY OF PRICE FLOORS AND PRICE CEILINGS**
Learning Objectives

- Explain how price floors and price ceilings can be inefficient

We demonstrated that market equilibrium maximizes social surplus; thus, the equilibrium quantity is the most efficient quantity of output for society. The imposition of a price floor or a price ceiling will prevent a market from adjusting to its equilibrium price and quantity, and thus will create an inefficient outcome. But there is an additional twist here. Along with creating inefficiency, price floors and ceilings also transfer some consumer surplus to producers, or some producer surplus to consumers.

In the following interactive graph (Figure 1), we can see this transfer in action:

Visit this page in your course online to view this presentation.

**Figure 1 (Interactive Graph). Inefficiency of Price Ceilings.**

As a result of the transfer of consumer surplus to producers (or producer surplus to consumers), two changes occur. First, an inefficient outcome occurs and the total surplus of society is reduced. The loss in social surplus that occurs when the economy produces at an inefficient quantity is called **deadweight loss**. In a very real sense, it is like money thrown away that benefits no one. In the last slide of the above activity, you can see the deadweight loss shown as the area $U + W$. When deadweight loss exists, it is possible for both consumer and producer surplus to be higher, in this case because the **price control** is blocking some suppliers and demanders from transactions that would be beneficial to both.

A second change from the price ceiling is that some of the producer surplus is transferred to consumers. After the price ceiling is imposed, the new consumer surplus is $T + V$, while the new producer surplus is $X$. In other words, the price ceiling transfers the area of surplus ($V$) from producers to consumers. Note that the gain to consumers is less than the loss to producers, which is just another way of seeing the deadweight loss.

Let's look at another interactive graph (Figure 2), this time with a price floor instead of a price ceiling:

Visit this page in your course online to view this presentation.

**Figure 2 (Interactive Graph). Inefficiency of Price Floors.**

The net effect of the price floor in the above activity is that the price floor causes the area $H$ to be transferred from consumer to producer surplus, but also causes a deadweight loss of $J + K$.

This analysis shows that a price ceiling, like a law establishing rent controls, will transfer some producer surplus to consumers—which helps to explain why consumers often favor them. Conversely, a price floor like a guarantee that farmers will receive a certain price for their crops will transfer some consumer surplus to producers, which explains why producers often favor them. However, both price floors and price ceilings block some transactions that buyers and sellers would have been willing to make, and creates deadweight loss. Removing such barriers, so that prices and quantities can adjust to their equilibrium level, will increase the economy’s social surplus.

**Summary**

Consumer surplus is the gap between the price that consumers are willing to pay, based on their preferences, and the market equilibrium price. Producer surplus is the gap between the price for which producers are willing to sell a product, based on their costs, and the market equilibrium price. Social surplus is the sum of consumer surplus and producer surplus. Social surplus is greater at the equilibrium quantity and price than it will be at any other quantity and price. Deadweight loss is loss in social surplus that occurs when the economy produces at an inefficient quantity.
Try It

This next question allow you to get as much practice as you need, as you can click the link at the top of the question (“Try another version of this question”) to get a new version of the question. Practice until you feel comfortable with this concept.
Visit this page in your course online to practice before taking the quiz.

Try It

This next question allow you to get as much practice as you need, as you can click the link at the top of the question (“Try another version of this question”) to get a new version of the question. Practice until you feel comfortable with this concept.
Visit this page in your course online to practice before taking the quiz.

Glossary

deadweight loss: the loss of economic value (i.e. social surplus) that occurs when a market operates at an inefficient quantity of output

LEARN BY DOING: CONSUMER AND PRODUCER SURPLUS

Rather than have you read more about consumer and producer surplus, we’d prefer to have you practice what you’ve learned so far and see for yourself if you understand it.

These questions allow you to get as much practice as you need, as you can click the link at the top of the first question (“Try another version of this question”) to get a new version of the questions. Practice until you feel comfortable doing these questions.

Try It

Visit this page in your course online to practice before taking the quiz.

Try It

Visit this page in your course online to practice before taking the quiz.
INTRODUCTION TO LABOR AND FINANCIAL MARKETS

What you’ll learn to do: examine ways that supply and demand apply to labor and financial markets

So far in this module, you have examined applications of supply and demand and how these concepts explain shortages, surpluses, and allocative efficiency. In this section, we will look at a couple more examples of supply and demand, and instead of focusing on markets for goods and services, we will see how these same principles apply to labor and financial markets.

LABOR AND FINANCIAL MARKETS

Learning Objectives

- Describe how the theories of supply & demand can be applied labor markets and financial markets
- Use the four-step process to predict how economic conditions cause a change in supply, demand, and equilibrium
The theories of supply and demand do not apply just to markets for goods. They apply to any market, even markets for labor and financial services. **Labor markets** are markets for employees or jobs. **Financial markets** are markets for saving or borrowing.

When we think about demand and supply curves in goods and services markets, it is easy to picture who the demanders and suppliers are: businesses produce the products and households buy them. Who are the demanders and suppliers in labor and financial service markets? In labor markets job seekers (individuals) are the suppliers of labor, while firms and other employers who hire labor are the demanders for labor. In financial markets, any individual or firm who saves contributes to the supply of money, and any who borrows (person, firm, or government) contributes to the demand for money.

As a college student, you most likely participate in both labor and financial markets. Employment is a fact of life for most college students: in 2011, according to the BLS, 52% of undergraduates worked part time and another 20% worked full time. Most college students are also heavily involved in financial markets, primarily as borrowers. Among full-time students, about half take out a loan to help finance their education each year, and those loans average about $6,000 per year. Many students also borrow for other expenses, like purchasing a car. We can analyze labor markets and financial markets with the same tools we use to analyze demand and supply in the goods markets. Let’s take a look at a few examples.

**Supply and Demand in Labor Markets**

Economic events can change the equilibrium salary (or wage) and quantity of labor. Consider how the wave of new information technologies, like computer and telecommunications networks, has affected low-skill and high-skill workers in the U.S. economy. From the perspective of employers who demand labor, these new technologies are often a substitute for low-skill laborers like file clerks who used to keep file cabinets full of paper records of transactions. However, the same new technologies are a complement to high-skill workers like managers, who benefit from the technological advances by being able to monitor more information, communicate more easily, and juggle a wider array of responsibilities. So, how will the new technologies affect the wages of high-skill and low-skill workers? For this question, let’s again use the four-step process of analyzing how shifts in supply or demand affect a market.

**Technology and Wage Inequality: The Four-Step Process**

**Step 1.** What did the markets for low-skill labor and high-skill labor look like before the arrival of the new technologies?

In Figure 1(a) and Figure 1(b), $S_0$ is the original supply curve for labor and $D_0$ is the original demand curve for labor in each market. In each graph, the original point of equilibrium, $E_0$, occurs at the price $W_0$ and the quantity $Q_0$.

![Figure 1. (a) The demand for low-skill labor shifts to the left when technology can do the job previously done by these workers. (b) New technologies can also increase the demand for high-skill labor in fields such as information technology and network administration.](image)
Step 2. Does the new technology affect the supply of labor from households or the demand for labor from firms?

Answer

The technology change described here affects demand for labor by firms that hire workers.

Step 3. Will the new technology increase or decrease demand?

Answer

Based on the description earlier, as the substitute for low-skill labor becomes available, demand for low-skill labor will shift to the left, from \( D_0 \) to \( D_1 \). As the technology complement for high-skill labor becomes cheaper, demand for high-skill labor will shift to the right, from \( D_0 \) to \( D_1 \).

Step 4. Compare the new equilibrium price and quantity to the original equilibrium price.

Answer

The new equilibrium for low-skill labor, shown as point \( E_1 \) with price \( W_1 \) and quantity \( Q_1 \), has a lower wage and quantity hired than the original equilibrium, \( E_0 \). The new equilibrium for high-skill labor, shown as point \( E_1 \) with price \( W_1 \) and quantity \( Q_1 \), has a higher wage and quantity hired than the original equilibrium (\( E_0 \)).

So, the demand and supply model predicts that the new computer and communications technologies will raise the pay of high-skill workers but reduce the pay of low-skill workers. Indeed, from the 1970s to the mid-2000s, the wage gap widened between high-skill and low-skill labor. According to the National Center for Education Statistics, in 1980, for example, a college graduate earned about 30% more than a high school graduate with comparable job experience, but by 2012, a college graduate earned about 60% more than an otherwise comparable high school graduate. Many economists believe that the trend toward greater wage inequality across the U.S. economy was primarily caused by the new technologies.

Supply and Demand in Financial Markets

Now let’s examine how the theories of supply and demand also affect financial markets. Imagine that the U.S. economy became viewed as a less desirable place for foreign investors to put their money because of fears about the growth of the U.S. public debt. Using the four-step process for analyzing how changes in supply and demand affect equilibrium outcomes, how would increased U.S. public debt affect the equilibrium price and quantity for capital in U.S. financial markets?

THE EFFECT OF GROWING U.S. DEBT: The Four-Step Process

Step 1. Draw a diagram showing demand and supply for financial capital that represents the original scenario in which foreign investors are pouring money into the U.S. economy.

Figure 2 shows a demand curve, \( D \), and a supply curve, \( S \), where the supply of capital includes the funds arriving from foreign investors. The original equilibrium \( E_0 \) occurs at interest rate \( R_0 \) and quantity of financial investment \( Q_0 \).
Step 2. Will the diminished confidence in the U.S. economy as a place to invest affect demand or supply of financial capital?

Answer

Yes, it will affect supply. Many foreign investors look to the U.S. financial markets to store their money in safe financial vehicles with low risk and stable returns. As the U.S. debt increases, debt servicing will increase—that is, more current income will be used to pay the interest rate on past debt. Increasing U.S. debt also means that businesses may have to pay higher interest rates to borrow money, because business is now competing with the government for financial resources.

Step 3. Will supply increase or decrease? When the enthusiasm of foreign investors’ for investing their money in the U.S. economy diminishes, the supply of financial capital shifts to the left. Figure 3 shows the supply curve shift from $S_0$ to $S_1$.

Step 4. Compare the new equilibrium price and quantity to the original equilibrium price.
Answer

Foreign investors’ diminished enthusiasm leads to a new equilibrium, \( E_1 \), which occurs at the higher interest rate, \( R_1 \), and the lower quantity of financial investment, \( Q_1 \).

The economy has experienced an enormous inflow of foreign capital. According to the U.S. Bureau of Economic Analysis, by the third quarter of 2014, U.S. investors had accumulated $24.6 trillion of foreign assets, but foreign investors owned a total of $30.8 trillion of U.S. assets. If foreign investors were to pull their money out of the U.S. economy and invest elsewhere in the world, the result could be a significantly lower quantity of financial investment in the United States, available only at a higher interest rate. This reduced inflow of foreign financial investment could impose hardship on U.S. consumers and firms interested in borrowing.

In a modern, developed economy, financial capital often moves invisibly through electronic transfers between one bank account and another. Yet these flows of funds can be analyzed with the same tools of demand and supply as markets for goods or labor.

Try It

Visit this page in your course online to view this presentation.

Glossary

- **financial markets**: supply and demand for financial services; i.e. saving & borrowing
- **labor markets**: supply and demand for jobs

PUTTING IT TOGETHER: APPLICATIONS OF SUPPLY AND DEMAND

This module showed that the theories of supply and demand can be applied to a variety of real world issues. Market outcomes can be evaluated based on the amount of net value created for society, which can be measured by consumer, producer and social surplus. Price ceilings and price floors result in deadweight loss—the loss of economic value caused by operating at an inefficient quantity of output.

Let’s return to the example of the minimum wage. Careful analysis shows that imposition of, or increases in the minimum wage have significant distributional effects. In other words, there are winners and losers from the policy. The winners are workers who continue to have a job, but are now paid a higher salary. The losers are businesses who have to pay more for their employees. This increase in production costs will be passed on, in part to consumers who will end up paying higher prices for the businesses’ products. So consumers lose also. The big losers, though, are the people who had jobs at the lower wage, but lose them when the minimum wage is increased. Which employees are most likely to lose their jobs, the most experienced and skilled, or the least experienced and skilled? Don’t forget that ultimately what matters is the size of these effects. These are the technical details that policy analysts will look at before making any recommendations to decision makers.
Consider Groupon, a website which offers significant discounts on purchases at businesses people frequently use. It’s not unusual to obtain 50% off the normal price. Why do customers like Groupon? Because it increases the consumer surplus they obtain on purchases.

Why do businesses offer Groupon campaigns? Part of it is advertising, to attract customers who aren’t familiar with those businesses. Some businesses offer regular Groupon deals. They must be doing this to increase their producer surplus (i.e., profit). This is likely part of a larger strategy, called price discrimination, which you will learn more about when you study the theory of the firm. For now, it is enough to understand that Groupon campaigns enhance producer surplus.

Since both consumer surplus and producer surplus increase, we can say that total economic (or social) surplus has increased. This is just another way of saying that transactions benefit both parties, or as economists would say, this is a more efficient outcome for society. Computing the additional consumer and producer surplus tells us by how much economic surplus has increased.

![Groupon Gift Card](image)
Why learn about elasticity?

Imagine going to your favorite coffee shop and having the waiter inform you that the pricing has changed. Instead of $3 for a cup of coffee, you will now be charged $2 for coffee, $1 for creamer, and $1 for your choice of sweetener. If you pay your usual $3 for a cup of coffee, you must choose between creamer and sweetener. If you want both, you now face an extra charge of $1. Sound absurd? Well, that’s the situation Netflix customers found themselves in—facing a 60 percent price hike to retain the same service.

In early 2011, Netflix consumers paid about $10 a month for a package consisting of streaming video and DVD rentals. In July 2011, the company announced a packaging change. Customers wishing to retain both streaming video and DVD rental would be charged $15.98 per month, a price increase of about 60 percent. How would customers of the fourteen-year-old firm react? Would they abandon Netflix? Would the ease of access to other venues make a difference in how consumers responded to the Netflix price change? In this module, the answers to these questions—about the change in quantity with respect to a change in price—will be explored through a concept economists call elasticity.

Elasticity measures the behavioral response of economic agents in a given situation. Here are some examples:

- If a business raises its prices, will that have a large or small impact on demand?
- If you get a pay raise, how much more will you spend on food, clothing or entertainment?
- If hot dogs go on sale at the grocery store, how much additional mustard will consumers purchase?
- If the local Italian restaurant puts their pizza on sale, will the additional number of pizzas sold offset the discount on each item? In other words, will their sales revenues for pizza go up or down?

These are important real-world questions that we’ll study in this module.
Also, before we get into the details: it can be easy to get hung up on the math of elasticity calculations. Learning to do these calculations is an important part of applying the elasticity principle, but the math will seem more intuitive if you master concept first: understanding what elasticity means in a particular context will help you see what you’re trying to calculate.

INTRODUCTION TO ELASTICITY

What you'll learn to do: explain the concept of elasticity

Elasticity is an economics concept that measures the responsiveness of one variable to changes in another variable. For example, if you raise the price of your product, how will that affect your sales numbers? The variables in this question are price and sales numbers. Elasticity explains how much one variable, say sales numbers, will change in response to another variable, like the price of the product.

Mastering this concept resembles learning to ride a bike: it's tough at first, but when you get it, you won’t forget. A rookie mistake is learning the calculations of elasticity but failing to grasp the idea. Make sure you don’t do this! First take time to understand the concepts—then the calculations can be used simply to explain them in a numerical way.

ELASTICITY OF DEMAND
Learning Objectives

- Describe and give examples of elasticity

Figure 1. Mr. Fantastic is elastic.

Think about the word elastic. It suggests that an item can be stretched. In economics, when we talk about elasticity, we’re referring to how much something will stretch or change in response to another variable. Consider a rubber band, a leather strap, and a steel ring. If you pull on two sides of a rubber band (or Mr. Fantastic), the force will cause it to stretch a lot. If you use the same amount of force to pull on the ends of a leather strap, it will stretch somewhat, but not as much as the rubber band. If you pull on either side of a steel ring, applying the same amount of force, it probably won’t stretch at all (unless you’re very strong). Each of these materials (the rubber band, the leather strap, and the steel ring) displays a different amount of elasticity in response to being pulled, and all three fall somewhere on a continuum from very stretchy (elastic) to barely stretchy (inelastic).

There are different kinds of economic elasticity—for example, price elasticity of demand, price elasticity of supply, income elasticity of demand, and cross-price elasticity of demand—but the underlying property is always the same: how responsive or sensitive one thing is to a change in another thing.

Try It

Visit this page in your course online to view this presentation.

Elastic and Inelastic Demand

Let’s think about elasticity in the context of price and quantity demanded. While the law of demand does tell us that more of a good will be bought at a lower price, it does not tell us how much the quantity demanded will increase because of the price change. For example, if a store owner raises prices, she can expect that the quantity demanded will drop, but she might not know how sensitive customers will be to the change. How many people will buy her products despite the price increase and how many people will be driven away?

If a small change in price creates a large change in the quantity demanded, then we would say that the demand is very elastic—that is, the demand is very sensitive to a change in price. If, on the other hand, a large change in price results in a very small change in demand in the quantity demanded, then we would say the demand is inelastic. As we will see later, elastic and inelastic are relative concepts. Here’s a way to keep this straight: demand is inelastic when consumers are insensitive to changes in price.
Consider the example of cigarette taxes and smoking rates—a classic example of inelastic demand. Cigarettes are taxed at both the state and federal level. As you might expect, the greater the amount of the tax increase, the fewer cigarettes are bought and consumed. While the taxes are somewhat of a deterrent, demand doesn’t decrease as much as the price increase, though. We can say, then, that the demand for cigarettes is relatively inelastic.

You might think that elasticity isn’t an important consideration when it comes to the price of cigarettes. Surely any reduction in the demand for cigarettes would be a good thing, right? Does it really matter whether whether the demand is elastic or inelastic? It does. The reason is that taxes on cigarettes serve two purposes: to raise tax revenue for government and to discourage smoking. On one hand, if a higher cigarette tax discourages consumption by quite a lot—meaning a very large reduction in cigarette sales—then the cigarette tax on each pack will not raise much revenue for the government. On the other hand, a higher cigarette tax that does not discourage consumption by much will actually raise more tax revenue for the government (but not have much impact on smoking rates). Thus, when Congress tries to calculate the effects of altering its cigarette tax, it must analyze how much the tax affects the quantity of cigarettes consumed. In other words, understanding the elasticity of cigarette demand is key to measuring the impact of taxes on government revenue AND public health.

Try It
Visit this page in your course online to view this presentation.

This issue reaches beyond governments and taxes; every firm faces a similar challenge. Every time a firm considers raising the price that it charges, it needs to know how much a price increase will reduce the quantity of its product that is demanded. Conversely, when a firm puts its products on sale, it wants assurance that the lower price will lead to a significantly higher quantity demanded.

GLossary

**elastic demand**: a high responsiveness of quantity demanded or supplied to changes in price

**elasticity**: an economics concept that measures responsiveness of one variable to changes in another variable

**inelastic demand**: a low responsiveness by consumers to price changes
Now that you have a general idea of what elasticity is, let’s consider some of the factors that can help us predict whether demand for a product is more or less elastic. The following are important considerations:

- **Substitutes**: Price elasticity of demand is fundamentally about substitutes. If it’s easy to find a substitute product when the price of a product increases, the demand will be more elastic. If there are few or no alternatives, demand will be less elastic.

- **Necessities vs. luxuries**: A necessity is something you absolutely must have, almost regardless of the price. A luxury is something that would be nice to have, but it’s not absolutely necessary. Consider the elasticity of demand for cookies. A buyer may enjoy a cookie, but it doesn’t fulfill a critical need the way a snow shovel after a blizzard or a life-saving drug does. In general, the greater the necessity of the product, the less elastic, or more inelastic, the demand will be, because substitutes are limited. The more luxurious the product is, the more elastic demand will be.

- **Share of the consumer’s budget**: If a product takes up a large share of a consumer’s budget, even a small percentage increase in price may make it prohibitively expensive to many buyers. Take rental housing that’s located close to downtown. Such housing might cost half of one’s budget. A small percentage increase in rent could cause renters to relocate to cheaper housing in the suburbs, rather than reduce their spending on food, utilities, and other necessities. Therefore the larger the share of an item in one’s budget, the more price elastic demand is likely to be. By contrast, suppose the local grocery store increased the price of toothpicks by 50 percent. Since toothpicks represent such a small part of a consumer’s budget, even a significant increase in price is likely to have only a small effect on demand. Thus, the smaller the share of an item in one’s budget, the more price inelastic demand is likely to be.

- **Short run versus long run**: Price elasticity of demand is usually lower in the short run, before consumers have much time to react, than in the long run, when they have greater opportunity to find substitute goods. Thus, demand is more price elastic in the long run than in the short run.

- **Competitive dynamics**: Goods that can only be produced by one supplier generally have inelastic demand, while products that exist in a competitive marketplace have elastic demand. This is because a competitive marketplace offers more options for the buyer.

---

**Try It**

Visit this page in your course online to check your understanding.

With these considerations in mind, take a moment to see if you can figure out which of the following products have elastic demand and which have inelastic demand. It may be helpful to remember that when the buyer is insensitive to price, demand is inelastic.

- Gasoline
- College textbooks
- Coffee
- Airline tickets
- Concert tickets
- Soft drinks
- Medical procedures
<table>
<thead>
<tr>
<th>Inelastic Demand</th>
<th>Elastic Demand</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Gasoline</strong></td>
<td><strong>Gas from a Particular Station</strong></td>
</tr>
<tr>
<td>- The demand for gasoline generally is fairly inelastic, especially in the short run. Car travel</td>
<td>- The demand for gasoline from any single gas station, or chain of gas stations, is highly elastic. Buyers can choose between comparable products based on price. There are often many stations in a small geographic area that are equally convenient.</td>
</tr>
<tr>
<td>requires gasoline. The substitutes for car travel offer less convenience and control. Much car travel is necessary for people to move between activities and can’t be reduced to save money. In the long run, though, more options are available, such as purchasing a more fuel-efficient car or choosing a job that is closer to where you work.</td>
<td></td>
</tr>
<tr>
<td><strong>Traditional Textbooks</strong></td>
<td><strong>New Textbook Distribution Channels</strong></td>
</tr>
<tr>
<td>- Generally an instructor assigns a textbook to the student, and the student who wants access to the</td>
<td>- Increasingly, students have new options to buy the same textbooks from different distribution channels at different price points. These include textbook rentals and digital versions of the text. The introduction of new distribution channels is increasing options for buyers and having an impact on the price elasticity for publishers.</td>
</tr>
<tr>
<td>learning materials must buy it, regardless of the price level. Because the student can’t easily identify another textbook or resource that will ensure the same content and grade for the class, he has no substitutes and must buy the book at any price (or opt not to buy it at all).</td>
<td></td>
</tr>
<tr>
<td><strong>Specialty Coffee Drinks</strong></td>
<td><strong>Black Coffee</strong></td>
</tr>
<tr>
<td>- Many coffee shops have developed branded drinks and specialized experiences in order to reduce substitutes and build customer loyalty. While black coffee is available almost universally, there are few substitutes for a Starbucks Java Chip Frappuccino. Demand for such products is more inelastic.</td>
<td>- Coffee is generally widely available at a level of quality that meets the needs of most buyers. The combination of a low price, relative to the buyer’s spending power, and the fact that the product is sold by many different suppliers in a competitive market, make the demand highly elastic.</td>
</tr>
<tr>
<td><strong>Concert Tickets</strong></td>
<td><strong>Airline Tickets</strong></td>
</tr>
<tr>
<td>- Only Taylor Swift can offer a Taylor Swift concert. She holds a monopoly on the creation and delivery</td>
<td>- Airline tickets are sold in a fiercely competitive market. Buyers can easily compare prices, and buyers experience the services provided by competitors as being very similar. Buyers can often choose not to travel if the cost is too high or substitute travel by car or train.</td>
</tr>
<tr>
<td>of that experience. There is no substitute, and loyal fans are willing to pay for the experience. Because it is a scarce resource and the delivery is tightly controlled by a single provider, access to concerts has inelastic demand.</td>
<td></td>
</tr>
<tr>
<td>Inelastic Demand</td>
<td>Elastic Demand</td>
</tr>
<tr>
<td>------------------</td>
<td>---------------</td>
</tr>
<tr>
<td><strong>Medical Procedures</strong></td>
<td><strong>Soft Drinks</strong></td>
</tr>
<tr>
<td>- Essential medical procedures have inelastic demand. The patient will pay what she can or what she must. In general, products that significantly affect health and well-being have inelastic demand.</td>
<td>- Soft drinks and many other nonessential items have highly elastic demand. There is competition among every brand and type of soda, and there are many substitutes for the entire category of soft drinks.</td>
</tr>
</tbody>
</table>

**Try It**

Visit this page in your course online to check your understanding.

**Glossary**

- **elastic demand**: a high responsiveness of quantity demanded or supplied to changes in price
- **inelastic demand**: a low responsiveness by consumers to price changes

**INTRODUCTION TO CALCULATING PRICE ELASTICITY**

What you’ll learn to do: explain the price elasticity of demand and price elasticity of supply, and compute both using the midpoint method
Remember, elasticity measures the responsiveness of one variable to changes in another variable. In the last section we looked at price elasticity of demand, or how much a change in price affects the quantity demanded. In this section we will dig deeper by learning how to calculate elasticity using the midpoint method. We'll also introduce the idea of elasticity of supply. Supply can also be elastic, since a change in price will influence the quantity supplied.

**CALCULATING ELASTICITY AND PERCENTAGE CHANGES**

**Learning Objectives**

- Mathematically differentiate between elastic, inelastic, and unitary elasticities of demand
- Calculate percentage changes, or growth rates
- Differentiate between the midpoint elasticity approach and the point elasticity approach in calculating elasticity

**Calculating Elasticity**

The formula for calculating elasticity is:
Let's look at the practical example mentioned earlier about cigarettes. Certain groups of cigarette smokers, such as teenage, minority, low-income, and casual smokers, are somewhat sensitive to changes in price: for every 10 percent increase in the price of a pack of cigarettes, the smoking rates drop about 7 percent. Plugging those numbers into the formula, we get

\[
\text{Price Elasticity of Demand} = \frac{\text{percent change in quantity}}{\text{percent change in price}}.
\]

Let's compute this:

\[
\text{Price Elasticity of Demand} = \frac{-7\%}{10\%} = -0.7
\]

Try It

Visit this page in your course online to check your understanding.

Inelastic, Elastic, and Unitary Demand

So what does the number -0.7 tell us about the elasticity of demand? The negative sign reflects the law of demand: at a higher price, the quantity demanded for cigarettes declines. All price elasticities of demand have a negative sign, so it's easiest to think about elasticity in absolute value, ignoring the negative sign. The fact that the result is less than one is more important than the negative sign. It tells us that the size of the quantity change is less than the size of the price change (i.e. the numerator in the elasticity formula is less than the denominator). This tells us that it would take a relatively large price change in order to cause a relatively small change in quantity demanded. In other words, consumer responsiveness to a change in price is relatively small. Therefore, when the elasticity is less than 1, we say that demand is inelastic.

The data above indicate that the demand for cigarettes by teenagers, minority, low income and casual smokers is relatively inelastic. Addicted adult smokers, though, are even less sensitive to changes in the price—most are willing to pay whatever it takes to support their smoking habit. We can say that their demand is even more inelastic than low income or casual smokers.

Different products have different price elasticities of demand. If the absolute value of the elasticity of some product is greater than one, it means that the change in the quantity demanded is greater than the change in price. This indicates a larger reaction to price change, which we describe as elastic. If the elasticity is equal to one, it means that the change in the quantity demanded is exactly equal to the change in price, so the demand response is exactly proportional to the change in price. We call this unitary elasticity, because unitary means one.

Watch It

Watch this video carefully to understand how to solve for elasticity and to see what the numerical values for elasticity mean when applied to economic situations.

Watch this video online: https://youtu.be/4oj_Inj6pXA

Try It

Visit this page in your course online to view this presentation.

Calculating Percentage Changes and Growth Rates

Before we dive deeper into solving for elasticity, let's first make sure we are comfortable calculating percentage changes, also known as a growth rates. The formula for computing a growth rate is straightforward:
Suppose that a job pays $10 per hour. At some point, the individual doing the job is given a $2-per-hour raise. The percentage change (or growth rate) in pay is

\[ \frac{\$2}{\$10} = 0.20 \text{ or } 20\%. \]

Now to solve for elasticity, we use the growth rate, or percentage change, of the quantity demanded as well as the percentage change in price in order to examine how these two variables are related. The price elasticity of demand is the ratio between the percentage change in the quantity demanded \((Q_d)\) and the corresponding percent change in price:

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

There are two general methods for calculating elasticities: the point elasticity approach and the midpoint (or arc) elasticity approach. Elasticity looks at the percentage change in quantity demanded divided by the percentage change in price, but which quantity and which price should be the denominator in the percentage calculation? The point approach uses the initial price and initial quantity to measure percent change. This makes the math easier, but the more accurate approach is the midpoint approach, which uses the average price and average quantity over the price and quantity change. (These are the price and quantity halfway between the initial point and the final point.) Let’s compare the two approaches. Suppose the quantity demanded of a product was 100 at one point on the demand curve, and then it moved to 103 at another point. The growth rate, or percentage change in quantity demanded, would be the change in quantity demanded divided by the average of the two quantities demanded:

\[
\frac{(103 - 100)}{2}.
\]

In other words, the growth rate:

\[
\frac{103 - 100}{(103 + 100)/2} = \frac{3}{101.5} = 0.0296 = 2.96\% \text{ growth}
\]

Note that if we used the point approach, the calculation would be:

\[
\frac{103 - 100}{100} = 3\% \text{ growth}
\]

This produces nearly the same result as the slightly more complicated midpoint method (3% vs. 2.96%). If you need a rough approximation, use the point method. If you need accuracy, use the midpoint method. Note: as the two points become closer together, the point elasticity becomes a closer approximation to the arc elasticity.

In this module you will often be asked to calculate the percentage change in the quantity. Keep in mind that this is same as the growth rate of the quantity. As you work through the course and find other applications for calculate growth rates, you will be well prepared.
Glossary

**elastic demand:** when the calculated elasticity of demand is greater than one, indicating a high responsiveness of quantity demanded or supplied to changes in price

**elastic supply:** when the calculated elasticity of either supply is greater than one, indicating a high responsiveness of quantity demanded or supplied to changes in price

**inelastic demand:** when the calculated elasticity of demand is less than one, indicating that a 1 percent increase in price paid by the consumer leads to less than a 1 percent change in purchases (and vice versa); this indicates a low responsiveness by consumers to price changes

**inelastic supply:** when the calculated elasticity of supply is less than one, indicating that a 1 percent increase in price paid to the firm will result in a less than 1 percent increase in production by the firm; this indicates a low responsiveness of the firm to price increases (and vice versa if prices drop)

**midpoint elasticity approach:** Most accurate approach to solving for elasticity in which the percent changes in quantity demanded and price are measured relative to the average quantity demanded and price; the initial quantity demand is subtracted from the new quantity demanded; then divided by the average of the two quantities demanded; similarly, the initial price is subtracted from the new price, then divided by the average of the two prices

**point elasticity approach:** approximate method for solving for elasticity in which the percent changes are measured relative to the initial quantity demanded and price; the initial quantity demanded is subtracted from the new quantity demanded, then divided by the initial quantity demanded; similarly, the initial price is subtracted from the new price, then divided by the initial price.

**unitary elasticity:** when the calculated elasticity is equal to one indicating that a change in the price of the good or service results in a proportional change in the quantity demanded or supplied

---

**CALCULATING PRICE ELASTICITIES USING THE MIDPOINT FORMULA**

Learning Objectives
Calculate price elasticity using the midpoint method
Differentiate between slope and elasticity

We have defined price elasticity of demand as the responsiveness of the quantity demanded to a change in the price. We also explained that price elasticity is defined as the percent change in quantity demanded divided by the percent change in price. In this section, you will get some practice computing the price elasticity of demand using the midpoint method.

The Midpoint Method

To calculate elasticity, we will use the average percentage change in both quantity and price. This is called the **midpoint method for elasticity** and is represented by the following equations:

\[
\text{percent change in quantity} = \frac{Q_2 - Q_1}{(Q_2 + Q_1) / 2} \times 100
\]

\[
\text{percent change in price} = \frac{P_2 - P_1}{(P_2 + P_1) / 2} \times 100
\]

The advantage of the midpoint method is that one obtains the same elasticity between two price points whether there is a price increase or decrease. This is because the formula uses the same base for both cases.

Exercise: Calculating the Price Elasticity of Demand

Let's calculate the elasticity from points B to A and from points G to H, shown in Figure 2, below.

**Figure 2. Calculating the Price Elasticity of Demand.** The price elasticity of demand is calculated as the percentage change in quantity divided by the percentage change in price.

**Elasticity from Point B to Point A**

**Step 1.** We know that \(\text{Price Elasticity of Demand} = \frac{\text{percent change in quantity}}{\text{percent change in price}}\)

**Step 2.** From the midpoint formula we know that
Step 3. We can use the values provided in the figure (as price decreases from $70 at point B to $60 at point A) in each equation:

\[
\begin{align*}
\text{percent change in quantity} &= \frac{Q_2 - Q_1}{(Q_2 + Q_1) \div 2} \times 100 \\
\text{percent change in price} &= \frac{P_2 - P_1}{(P_2 + P_1) \div 2} \times 100
\end{align*}
\]

Step 4. Then, those values can be used to determine the price elasticity of demand:

\[
\text{Price Elasticity of Demand} = \frac{\text{percent change in quantity}}{\text{percent change in price}} = \frac{6.9 \text{ percent}}{-15.5 \text{ percent}} = -0.45
\]

The elasticity of demand between these two points is 0.45, which is an amount smaller than 1. That means that the demand in this interval is inelastic.

Remember: price elasticities of demand are always negative, since price and quantity demanded always move in opposite directions (on the demand curve). As you’ll recall, according to the law of demand, price and quantity demanded are inversely related. By convention, we always talk about elasticities as positive numbers, however. So, mathematically, we take the absolute value of the result. For example, -0.45 would interpreted as 0.45.

This means that, along the demand curve between points B and A, if the price changes by 1%, the quantity demanded will change by 0.45%. A change in the price will result in a smaller percentage change in the quantity demanded. For example, a 10% increase in the price will result in only a 4.5% decrease in quantity demanded. A 10% decrease in the price will result in only a 4.5% increase in the quantity demanded.

Note also that a larger (negative) number means demand is more elastic, so that if price elasticity of demand were -0.75, the quantity demanded would change by a greater percentage than when the elasticity was -0.45.

Exercise: Elasticity Of Demand from Point G to Point H

Calculate the price elasticity of demand using the data in Figure 2 for an increase in price from G to H. Does the elasticity increase or decrease as we move up the demand curve?

Step 1. We know that Price Elasticity of Demand = \( \frac{\text{percent change in quantity}}{\text{percent change in price}} \)

Step 2. From the midpoint formula we know that

\[
\begin{align*}
\text{percent change in quantity} &= \frac{Q_2 - Q_1}{(Q_2 + Q_1) \div 2} \times 100 \\
\text{percent change in price} &= \frac{P_2 - P_1}{(P_2 + P_1) \div 2} \times 100
\end{align*}
\]

Step 3. We can use the values provided in the figure in each equation:

\[
\begin{align*}
\text{percent change in quantity} &= \frac{1,600 - 1,800}{(1,600 + 1,800) \div 2} \times 100 = \frac{-200}{1,700} \times 100 = -11.76 \\
\text{percent change in price} &= \frac{130 - 120}{(130 + 120) \div 2} \times 100 = \frac{10}{125} \times 100 = 8.0
\end{align*}
\]

Step 4. Then, those values can be used to determine the price elasticity of demand:

\[
\text{Price Elasticity of Demand} = \frac{\text{percent change in quantity}}{\text{percent change in price}} = \frac{-11.76}{8} = 1.45
\]

The elasticity of demand from G to H is 1.45. The magnitude of the elasticity has increased (in absolute value) as we moved up along the demand curve from points A to B. Recall that the elasticity between those two points is 0.45. Demand is inelastic between points A and B and elastic between points G and H. This shows us that price elasticity of demand changes at different points along a straight-line demand curve.
Let's pause and think about why the elasticity is different over different parts of the demand curve. When price elasticity of demand is greater (as between points G and H), it means that there is a larger impact on demand as price changes. That is, when the price is higher, buyers are more sensitive to additional price increases. Logically, that makes sense.

**Elasticity Is Not Slope**

It's a common mistake to confuse the slope of either the supply or demand curve with its elasticity. The slope is the rate of change in units along the curve, or the rise/run (change in y over the change in x). For example, in Figure 2 above, for each point shown on the demand curve, price drops by $10 and the number of units demanded increases by 200. So the slope is $-10/200$ along the entire demand curve, and it doesn't change. The price elasticity, however, changes along the curve. Elasticity between points B and A was 0.45 and increased to 1.47 between points G and H. Elasticity is the percentage change—which is a different calculation from the slope, and it has a different meaning.

When we are at the upper end of a demand curve, where price is high and the quantity demanded is low, a small change in the quantity demanded—even by, say, one unit—is pretty big in percentage terms. A change in price of, say, a dollar, is going to be much less important in percentage terms than it will be at the bottom of the demand curve. Likewise, at the bottom of the demand curve, that one unit change when the quantity demanded is high will be small as a percentage. So, at one end of the demand curve, where we have a large percentage change in quantity demanded over a small percentage change in price, the elasticity value will be high—demand will be relatively elastic. Even with the same change in the price and the same change in the quantity demanded, at the other end of the demand curve the quantity is much higher, and the price is much lower, so the percentage change in quantity demanded is smaller and the percentage change in price is much higher. See Figure 3, below:

![Figure 3. Elasticity changes along the demand curve.](image)

At the bottom of the curve we have a small numerator over a large denominator, so the elasticity measure will be much lower, or inelastic. As we move along the demand curve, the values for quantity and price go up or down, depending on
which way we are moving, so the percentages for, say, a $1 difference in price or a one-unit difference in quantity, will change as well, which means the ratios of those percentages will change, too.

**Try It**

Visit this page in your course online to view this presentation.

**Try It**

These next questions allow you to get as much practice as you need, as you can click the link at the top of the questions (“Try another version of these questions”) to get a new version of the questions. Practice until you feel comfortable with this concept.

Visit this page in your course online to practice before taking the quiz.

**Glossary**

**elasticity**: an economics concept that measures responsiveness of one variable to changes in another variable

**midpoint method**: measures the average elasticity over some part of the demand (or supply) curve

**more elastic**: the calculated elasticity is greater in absolute value, meaning the quantity response is greater to the same change in price

**LEARN BY DOING: CALCULATING PRICE ELASTICITIES**

**Try It**

These questions allow you to get as much practice as you need, as you can click the link at the top of the first question (“Try another version of these questions”) to get a new set of questions. Practice until you feel comfortable doing the questions.

Visit this page in your course online to practice before taking the quiz.
CATEGORIES OF ELASTICITY

Learning Objectives

- Explain and compare the graphs for the following types of elasticities: elastic, inelastic, unitary, infinite, and zero

The language of elasticity can sometimes be confusing. We use the word elasticity to describe the property of responsiveness in economic variables. We also describe the responsiveness as (relatively) elastic or (relatively) inelastic. It gets worse. We can also describe elasticity as perfectly elastic or perfectly inelastic. How to we keep these different meanings understood? That is the purpose of this section.

We mentioned previously that elasticity measurements are divided into three main ranges: elastic, inelastic, and unitary, corresponding to different parts of a linear demand curve.

Demand is described as elastic when the computed elasticity is greater than 1, indicating a high responsiveness to changes in price. Computed elasticities that are less than 1 indicate low responsiveness to price changes and are described as inelastic demand. Unitary elasticities indicate proportional responsiveness of demand. In other words, the percent change in quantity demanded is equal to the percent change in price, so the elasticity equals 1. These ranges are summarized in Table 1, below.

<table>
<thead>
<tr>
<th>If . . .</th>
<th>Then . . .</th>
<th>And It’s Called . . .</th>
</tr>
</thead>
<tbody>
<tr>
<td>% change in quantity &gt; % change in price</td>
<td>Computed Elasticity &gt; 1</td>
<td>Elastic</td>
</tr>
<tr>
<td>% change in quantity = % change in price</td>
<td>Computed Elasticity = 1</td>
<td>Unitary</td>
</tr>
<tr>
<td>% change in quantity &lt; % change in price</td>
<td>Computed Elasticity &lt; 1</td>
<td>Inelastic</td>
</tr>
</tbody>
</table>

It is important to note that both elastic and inelastic are relative terms, as shown in Figure 1, below. As one moves down the demand curve from top left to bottom right, the measured elasticity is much greater than one (very elastic), then just greater than one (somewhat elastic), then equal to one (unitary elastic, then less than one (somewhat inelastic), and finally much less than one (very inelastic). Note that the epsilon symbol, ε, is often used to represent elasticity.
Polar Cases of Elasticity

There are also two extreme cases of elasticity: when computed elasticity equals zero and when it’s infinite. We will describe each case.

A **perfectly** (or **infinitely** elastic) demand curve refers to the extreme case in which the quantity demanded (Qd) increases by an infinite amount in response to any decrease in price at all. Similarly, quantity demanded drops to zero for any increase in the price. A perfectly elastic demand curve is horizontal, as shown in Figure 2, below. While it’s difficult to think of real world example of infinite elasticity, it will be important when we study perfectly competitive markets. It’s a situation where consumers are extremely sensitive to changes in price. Say, for example, if the price of cruises to the Caribbean decreased, everyone would buy tickets (i.e., quantity demanded would increase to infinity), or when the price of cruises to the Caribbean increased, not a single person would be on the boat (i.e., quantity demanded would decrease to zero). Perfectly elastic demand is an “all or nothing” thing!

*Figure 2. Infinite Elasticity.* This shows a perfectly elastic demand curve. The horizontal line shows that an infinite quantity will be demanded at a specific price. The quantity demanded is extremely responsive to price changes, moving from zero for prices close to P to infinite when prices reach P.

While **perfectly inelastic demand** is an extreme case, necessities with no close substitutes are likely to have highly inelastic demand curves. This is the case with life-saving prescription drugs, for example. Consider a person with kidney...
failure who needs insulin to stay alive. A specific quantity of insulin is prescribed to the patient. If the price of insulin decreases, the patient can’t stock up and save it for the future. If the price of insulin increases, the patient will continue to purchase the same quantity needed to stay alive. Perfectly inelastic demand means that quantity demanded remains the same when price increases or decreases. Consumers are completely unresponsive to changes in price.

Final note: even though perfectly elastic and perfectly inelastic curves correspond to horizontal and vertical curves, remember that, in general, elasticity is not the same as the slope.

Watch It

Watch this video to see graphed examples of perfectly inelastic, relatively inelastic, unit elastic, relatively elastic, and perfectly elastic demand.
Watch this video online: https://youtu.be/HHcbllxiAAk

Try It

Visit this page in your course online to view this presentation.

Try It

These questions allow you to get as much practice as you need, as you can click the link at the top of the first question (“Try another version of these questions”) to get a new set of questions. Practice until you feel comfortable doing the questions.
Visit this page in your course online to practice before taking the quiz.

Glossary

**perfectly (or infinitely) elastic**: the extremely elastic situation of demand or supply where quantity changes by an infinite amount in response to any change in price; horizontal in appearance

**perfectly inelastic**: the highly inelastic case of demand in which a percentage change in price, no matter how large, results in zero change in the quantity; thus, the price elasticity of demand is zero; vertical in appearance

**(relatively) elastic**: the percentage change in quantity demanded is greater than the percentage change in price; measured price elasticity of demand is greater than one (in absolute value)
(relatively) inelastic: the percentage change in quantity demanded is less than the percentage change in price; measure price elasticity of demand is less than one (in absolute value)

unitary elastic: when a given percent price change in price leads to an equal percentage change in quantity demanded

PRICE ELASTICITY OF SUPPLY

Learning Objectives

- Calculate the price elasticity of supply

Calculating the Price Elasticity of Supply

The price elasticity of supply measures how much quantity supplied changes in response to a change in the price. The calculations and interpretations are analogous to those we explained above for the price elasticity of demand. The only difference is we are looking at how producers respond to a change in the price instead of how consumers respond.

Price elasticity of supply is the percentage change in the quantity of a good or service supplied divided by the percentage change in the price. Since this elasticity is measured along the supply curve, the law of supply holds, and thus price elasticities of supply are always positive numbers. We describe supply elasticities as elastic, unitary elastic and inelastic, depending on whether the measured elasticity is greater than, equal to, or less than one.

Exercise: Elasticity of Supply from Point A to Point B

Assume that an apartment rents for $650 per month and at that price 10,000 units are offered for rent, as shown in Figure 2, below. When the price increases to $700 per month, 13,000 units are offered for rent. By what percentage does apartment supply increase? What is the price sensitivity?
Figure 2. Price Elasticity of Supply. The price elasticity of supply is calculated as the percentage change in quantity divided by the percentage change in price.

Step 1. We know that

\[
\text{Price Elasticity of Supply} = \frac{\text{percent change in quantity}}{\text{percent change in price}}
\]

Step 2. From the midpoint method we know that

\[
\text{percent change in quantity} = \frac{Q_2 - Q_1}{(Q_2 + Q_1) \div 2} \times 100
\]

\[
\text{percent change in price} = \frac{P_2 - P_1}{(P_2 + P_1) \div 2} \times 100
\]

Step 3. We can use the values provided in the figure in each equation:

\[
\text{percent change in quantity} = \frac{13,000 - 10,000}{(13,000 + 10,000) \div 2} \times 100 = \frac{3,000}{11,500} \times 100 = 26.1
\]

\[
\text{percent change in price} = \frac{750 - 600}{(750 + 600) \div 2} \times 100 = \frac{50}{675} \times 100 = 7.4
\]

Step 4. Then, those values can be used to determine the price elasticity of demand:

\[
\text{Price Elasticity of Supply} = \frac{26.1 \text{ percent}}{7.4 \text{ percent}} = 3.53
\]

Again, as with the elasticity of demand, the elasticity of supply is not followed by any units. Elasticity is a ratio of one percentage change to another percentage change—nothing more—and is read as an absolute value. In this case, a 1% rise in price causes an increase in quantity supplied of 3.5%. Since 3.5 is greater than 1, this means that the percentage change in quantity supplied will be greater than a 1% price change. If you’re starting to wonder if the concept of slope fits into this calculation, read on for clarification.

Watch It

Watch this video to see a real-world application of price elasticity.

Watch this video online: https://youtu.be/XOC_nIQ5its
INTRODUCTION TO ELASTICITIES IN AREAS OTHER THAN PRICE

What you’ll learn to do: explain and calculate other elasticities using common economic variables
Remember, elasticity measures the responsiveness of one variable to changes in another variable. We have focused on how a change in price can impact other variables. Elasticity doesn't apply only to price, however. It can describe anything that affects demand or supply. For example, when consumer income varies, it can have an impact on demand. When we consider that impact, we are measuring the responsiveness of one variable (demand) to changes in another variable (consumer income). This is called the income elasticity of demand.

Likewise, if two goods are complements or substitutes, a change in demand for one can have an impact on the demand for the other. This is known as cross-price elasticity of demand. In this section, we'll elaborate on the idea of elasticity to see how it applies to other economic variables.

### INCOME ELASTICITY, CROSS-PRICE ELASTICITY & OTHER TYPES OF ELASTICITIES

#### Learning Objectives

- Calculate the income elasticity of demand
- Explain and calculate cross-price elasticity of demand
- Describe elasticity in labor and financial capital markets
The basic idea of elasticity—how a percentage change in one variable causes a percentage change in another variable—does not just apply to the responsiveness of supply and demand to changes in the price of a product. Recall that quantity demanded (Qd) depends on income, tastes and preferences, population, expectations about future prices, and the prices of related goods. Similarly, quantity supplied (Qs) depends on the cost of production, changes in weather (and natural conditions), new technologies, and government policies. Elasticity can, in principle, be measured for any determinant of supply and demand, not just the price.

### Income Elasticity of Demand

The **income elasticity of demand** is the percentage change in quantity demanded divided by the percentage change in income, as follows:

\[
\text{income elasticity of demand} = \frac{\text{percent change in quantity demanded}}{\text{percent change in income}}
\]

For most products, most of the time, the income elasticity of demand is positive: that is, a rise in income will cause an increase in the quantity demanded. This pattern is common enough that these goods are referred to as **normal goods**. However, for a few goods, an increase in income means that one might purchase less of the good; for example, those with a higher income might buy fewer hamburgers, because they are buying more steak instead, or those with a higher income might buy less cheap wine and more imported beer. When the income elasticity of demand is negative, the good is called an **inferior good**. The concepts of normal and inferior goods were introduced in the Supply and Demand module. A higher level of income for a normal good causes a demand curve to shift to the right for a normal good, which means that the income elasticity of demand is positive. How far the demand shifts depends on the income elasticity of demand. A higher income elasticity means a larger shift. However, for an inferior good—that is, when the income elasticity of demand is negative—a higher level of income would cause the demand curve for that good to shift to the left. Again, how much it shifts depends on how large the (negative) income elasticity is.

**Watch It**

Watch this video online: [https://youtu.be/a6AHaqlm7J4](https://youtu.be/a6AHaqlm7J4)

**Try It**
Cross-Price Elasticity of Demand

A change in the price of one good can shift the quantity demanded for another good. If the two goods are complements, like bread and peanut butter, then a drop in the price of one good will lead to an increase in the quantity demanded of the other good. However, if the two goods are substitutes, like plane tickets and train tickets, then a drop in the price of one good will cause people to substitute toward that good, and to reduce consumption of the other good. Cheaper plane tickets lead to fewer train tickets, and vice versa. The cross-price elasticity of demand puts some meat on the bones of these ideas. The term "cross-price" refers to the idea that the price of one good is affecting the quantity demanded of a different good. Specifically, the cross-price elasticity of demand is the percentage change in the quantity of good A that is demanded as a result of a percentage change in the price of good B, as follows:

\[
\text{cross-price elasticity of demand} = \frac{\text{percent change in } Q_d \text{ of good } A}{\text{percent change in price of good } B}
\]

 Substitute goods have positive cross-price elasticities of demand: if good A is a substitute for good B, like coffee and tea, then a higher price for B will mean a greater quantity of A consumed. Complement goods have negative cross-price elasticities: if good A is a complement for good B, like coffee and sugar, then a higher price for B will mean a lower quantity of A consumed.

Exercise: Calculating Cross-Price Elasticity of Demand

Let's practice calculating cross-price elasticity of demand by looking at two goods: widgits and sprockets. The initial price and quantity of widgets demanded is \((P_1 = 12, Q_1 = 8)\). The subsequent price and quantity is \((P_2 = 9, Q_2 = 10)\). This is all the information needed to compute the price elasticity of demand.

The price elasticity of demand is defined as follows:

\[
\text{Price Elasticity of Demand} = \frac{\text{percent change in quantity}}{\text{percent change in price}}
\]

From the midpoint formula, we know that:

\[
\text{percent change in quantity} = \frac{Q_2 - Q_1}{(Q_2 + Q_1) / 2} \times 100 = \frac{10 - 8}{(10 + 8) / 2} \times 100 = \frac{2}{9} \times 100 = 22.2
\]

And:

\[
\text{percent change in price} = \frac{P_2 - P_1}{(P_2 + P_1) / 2} \times 100 = \frac{9 - 12}{(9 + 12) / 2} \times 100 = \frac{-3}{10.5} \times 100 = -28.6
\]

Therefore:

\[
\text{Price Elasticity of Demand} = \frac{22.2 \text{ percent}}{-28.6 \text{ percent}} = -0.77
\]

Since the elasticity is less than 1 (in absolute value), we can say that the price elasticity of demand for widgets is in the inelastic range.

The cross-price elasticity of demand is computed similarly:

\[
\text{Cross-Price Elasticity of Demand} = \frac{\text{percent change in quantity of sprockets demanded}}{\text{percent change in price of widgets}}
\]

The initial quantity of sprockets demanded is 9 and the subsequent quantity demanded is 10 \((Q_1 = 9, Q_2 = 10)\). Using the midpoint formula, we can calculate the percent change in the quantity of sprockets demanded:

\[
\text{percent change in quantity} = \frac{Q_2 - Q_1}{(Q_2 + Q_1) / 2} \times 100 = \frac{10 - 9}{(10 + 9) / 2} \times 100 = \frac{1}{9.5} \times 100 = 10.5
\]

The percent change in the quantity of sprockets demanded is 10.5%.

The percent change in the price of widgets is the same as above, or -28.6%.

Therefore:

\[
\text{Cross-Price Elasticity of Demand} = \frac{10.5 \text{ percent}}{-28.6 \text{ percent}} = -0.37
\]
Because the cross-price elasticity is negative, we can conclude that widgets and sprockets are complementary goods. Intuitively, when the price of widgets goes down, consumers purchase more widgets. Because they’re purchasing more widgets, they purchase more sprockets.

Watch It

Watch this video to see an example of solving for the cross-price elasticity of demand.

Watch this video online: https://youtu.be/FgSSLAWq_nE

Try It

Visit this page in your course online to view this presentation.

Elasticity in Labor and Financial Capital Markets

The concept of elasticity applies to any market, not just markets for goods and services. In the labor market, for example, the wage elasticity of labor supply—that is, the percentage change in hours worked divided by the percentage change in wages—will determine the shape of the labor supply curve. The formula is as follows:

$$\text{elasticity of labor supply} = \frac{\text{percent change in quantity of labor supplied}}{\text{percent change in wage}}$$

The wage elasticity of labor supply for teenage workers is generally thought to be fairly elastic: that is, a certain percentage change in wages will lead to a larger percentage change in the quantity of hours worked. Conversely, the wage elasticity of labor supply for adult workers in their thirties and forties is thought to be fairly inelastic. When wages move up or down by a certain percentage amount, the quantity of hours that adults in their prime earning years are willing to supply changes but by a lesser percentage amount. In markets for financial capital, the interest elasticity of savings—that is, the percentage change in the quantity of savings divided by the percentage change in interest rates—will describe the shape of the supply curve for financial capital, as follows:

$$\text{interest elasticity of savings} = \frac{\text{percent change in quantity of financial savings}}{\text{percent change in interest rate}}$$

Sometimes laws are proposed that seek to increase the quantity of savings by offering tax breaks so that the return on savings is higher. Such a policy will increase the quantity if the supply curve for financial capital is elastic, because then a given percentage increase in the return to savings will cause a higher percentage increase in the quantity of savings. However, if the supply curve for financial capital is highly inelastic, then a percentage increase in the return to savings will cause only a small increase in the quantity of savings. The evidence on the supply curve of financial capital is controversial but, at least in the short run, the elasticity of savings with respect to the interest rate appears fairly inelastic.

Expanding the Concept of Elasticity

The elasticity concept does not even need to relate to a typical supply or demand curve at all. For example, imagine that you are studying whether the Internal Revenue Service should spend more money on auditing tax returns. The question can be framed in terms of the elasticity of tax collections with respect to spending on tax enforcement; that is, what is the percentage change in tax collections derived from a percentage change in spending on tax enforcement? With all of the elasticity concepts that have just been described, some of which are listed in Table 1, the possibility of confusion arises. When you hear the phrases “elasticity of demand” or “elasticity of supply,” they refer to the elasticity with respect to price. Sometimes, either to be extremely clear or because a wide variety of elasticities is being discussed, the elasticity of demand or the demand elasticity will be called the price elasticity of demand or the “elasticity of demand with respect to price.” Similarly, elasticity of supply or the supply elasticity is sometimes called, to avoid any possibility of confusion, the price elasticity of supply or “the elasticity of supply with respect to price.” But in whatever context
elasticity is invoked, the idea always refers to percentage change in one variable, almost always a price or money variable, and how it causes a percentage change in another variable, typically a quantity variable of some kind.

<table>
<thead>
<tr>
<th>Table 1. Formulas for Calculating Elasticity</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Elasticity Type</strong></td>
</tr>
<tr>
<td>Income elasticity of demand</td>
</tr>
<tr>
<td>Cross-price elasticity of demand</td>
</tr>
<tr>
<td>Wage elasticity of labor supply</td>
</tr>
<tr>
<td>Wage elasticity of labor demand</td>
</tr>
<tr>
<td>Interest rate elasticity of savings</td>
</tr>
<tr>
<td>Interest rate elasticity of borrowing</td>
</tr>
</tbody>
</table>

Try It

Visit this page in your course online to view this presentation.

Glossary

cross-price elasticity of demand: the percentage change in the quantity of good A that is demanded as a result of a percentage change in good B

interest elasticity of savings: the percentage change in the quantity of savings divided by the percentage change in interest rates

wage elasticity of labor supply: the percentage change in hours worked divided by the percentage change in wages
INTRODUCTION TO PRICE ELASTICITY AND TOTAL REVENUE

What you’ll learn to do: explain the relationship between a firm’s price elasticity of demand and total revenue

Price elasticity of demand describes how changes in the price for goods and the demand for those same goods relate. As those two variables interact, they can have an impact on a firm’s total revenue. Revenue is the amount of money a firm brings in from sales—i.e., the total number of units sold multiplied by the price per unit. Therefore, as the price or the quantity sold changes, those changes have a direct impact on revenue.

Businesses seek to maximize their profits, and price is one tool they have at their disposal to influence demand (and therefore sales). Picking the right price is tricky, though. What happens with a price increase? Will customers buy only a little less, such that the price increase raises revenues, or will they buy a lot less, such that the price increase lowers revenues? Might the company earn more if it lowers prices, or will that just lead to lower revenue per unit without stimulating new demand? These are critical questions for every business.

ELASTICITY AND TOTAL REVENUE

Learning Objectives
Total Revenue and Elasticity of Demand

Studying elasticities is useful for a number of reasons, pricing being the most important. Imagine that a band on tour is playing in an indoor arena with 15,000 seats. To keep this example simple, assume that the band keeps all the money from ticket sales. Assume further that the band pays the costs for its appearance, but that these costs, like travel, setting up the stage, and so on, are the same regardless of how many people are in the audience. Finally, assume that all the tickets have the same price. (The same insights apply if ticket prices are more expensive for some seats than for others, but the calculations become more complicated.) The band knows that it faces a downward-sloping demand curve; that is, if the band raises the price of tickets, it will sell fewer tickets. How should the band set the price for tickets to bring in the most total revenue, which in this example, because costs are fixed, will also mean the highest profits for the band? Should the band sell more tickets at a lower price or fewer tickets at a higher price?

The key concept in thinking about collecting the most revenue is the price elasticity of demand. Total revenue is price times the quantity of tickets sold (TR = P x Qd). Imagine that the band starts off thinking about a certain price, which will result in the sale of a certain quantity of tickets. The three possibilities are laid out in Table 1. If demand is elastic at that price level, then the band should cut the price, because the percentage drop in price will result in an even larger percentage increase in the quantity sold—thus raising total revenue. However, if demand is inelastic at that original quantity level, then the band should raise the price of tickets, because a certain percentage increase in price will result in a smaller percentage decrease in the quantity sold—and total revenue will rise. If demand has a unitary elasticity at that quantity, then a moderate percentage change in the price will be offset by an equal percentage change in quantity—so the band will earn the same revenue whether it (moderately) increases or decreases the price of tickets.

Table 1. Price Elasticity of Demand

<table>
<thead>
<tr>
<th>If demand is . . .</th>
<th>Then . . .</th>
<th>Therefore . . .</th>
</tr>
</thead>
</table>
| Elastic           | % change in Qd is greater than % change in P | • A given % rise in P will be more than offset by a larger % fall in Q so that total revenue (P times Q) falls.  
• A given % fall in P will be more than offset by a larger rise in Q so that total revenue (P times Q) rises. |
| Unitary           | % change in Qd is equal to % change in P | • A given % rise or fall in P will be exactly offset by an equal % fall in Q so that total revenue (P times Q) is unchanged. |
| Inelastic         | % change in Qd is less than % change in P | • A given % rise in P will cause a smaller % fall in Q so that total revenue (P times Q) rises.  
• A given % fall in P will cause a smaller % rise in Q so that total revenue (P times Q) falls. |

If demand is elastic at a given price level, then should a company cut its price, the percentage drop in price will result in an even larger percentage increase in the quantity sold—thus raising total revenue. However, if demand is inelastic at the original quantity level, then should the company raise its prices, the percentage increase in price will result in a smaller percentage decrease in the quantity sold—and total revenue will rise.

Let’s explore some specific examples. In both cases we will answer the following questions:

1. How much of an impact do we think a price change will have on demand?
2. How would we calculate the elasticity, and does it confirm our assumption?
3. What impact does the elasticity have on total revenue?


Example 1: The Student Parking Permit

How elastic is the demand for student parking passes at your institution?

The answer to that question likely varies based on the profile of your institution, but we are going to explore a particular example. Let's consider a community college campus where all of the students commute to class. Required courses are spread throughout the day and the evening, and most of the classes require classroom attendance (rather than online participation). There is a reasonable public transportation system with busses coming to and leaving campus from several lines, but the majority of students drive to campus. A student parking permit costs $40 per term. As the parking lots become increasingly congested, the college considers raising the price of the parking passes in hopes that it will encourage more students to carpool or to take the bus.

If the college increases the price of a parking permit from $40 to $48, how many fewer students will buy parking permits?

If you think that the change in price will cause many students to decide not to buy a permit, then you are suggesting that the demand is elastic—the students are quite sensitive to price changes. If you think that the change in price will not impact student permit purchases much, then you are suggesting that the demand is inelastic—student demand for permits is insensitive to price changes.

In this case, we can all argue that students are very sensitive to increases in costs in general, but the determining factor in their demand for parking permits is more likely to be the quality of alternative solutions. If the bus service does not allow students to travel between home, school, and work in a reasonable amount of time, many students will resort to buying a parking permit, even at the higher price. Because students don't generally have extra money, they may grumble about a price increase, but many will still have to pay.

Let's add some numbers and test our thinking. The college implements the proposed increase of $8, taking the new price to $48. Last year the college sold 12,800 student parking passes. This year, at the new price, the college sells 11,520 parking passes.

\[
\text{percent change in quantity} = \frac{11,520 - 12,800}{(11,520 + 12,800) \div 2} \times 100 = \frac{-1280}{12160} \times 100 = -10.53
\]

\[
\text{percent change in price} = \frac{48 - 40}{(48 + 40) \div 2} \times 100 = \frac{8}{44} \times 100 = 18.18
\]

\[
\text{Price Elasticity of Demand} = \frac{-10.53 \text{ percent}}{18.18 \text{ percent}} = -.58
\]

First, looking only at the percent change in quantity and the percent change in price we know that an 18% change in price will result in an 11% change in demand. In other words, a large change in price created a comparatively smaller change in demand. We can also see that the elasticity is 0.58. When the absolute value of the price elasticity is < 1, the demand is inelastic. In this example, student demand for parking permits is inelastic.

What impact does the price change have on the college and their goals for students? First, there are 1,280 fewer cars taking up parking places. If all of those students are using alternative transportation to get to school and this change has relieved parking-capacity issues, then the college may have achieved its goals. However, there’s more to the story: the price change also has an effect on the college’s revenue, as we can see below:

Year 1: 12,800 parking permits sold x $40 per permit = $512,000

Year 2: 11,520 parking permits sold x $48 per permit = $552,960

The college earned an additional $40,960 in revenue. Perhaps this can be used to expand parking or address other student transportation issues.

In this case, student demand for parking permits is inelastic. A significant change in price leads to a comparatively smaller change in demand. The result is lower sales of parking passes but more revenue.

Note: If you attend an institution that offers courses completely or largely online, the price elasticity for parking permits might be perfectly inelastic. Even if the institution gave away parking permits, students might not want them.
Example 2: Helen’s Cookies

Have you been at the counter of a convenience store and seen cookies for sale on the counter? In this example we are going to consider a baker, Helen, who bakes these cookies and sells them for $2 each. The cookies are sold in a convenience store, which has several options on the counter that customers can choose as a last-minute impulse buy. All of the impulse items range between $1 and $2 in price. In order to raise revenue, Helen decides to raise her price to $2.20.

If Helen increases the cookie price from $2.00 to $2.20—a 10% increase—will fewer customers buy cookies?

If you think that the change in price will cause many buyers to forego a cookie, then you are suggesting that the demand is elastic, or that the buyers are sensitive to price changes. If you think that the change in price will not impact sales much, then you are suggesting that the demand for cookies is inelastic, or insensitive to price changes.

Let’s assume that this price change does impact customer behavior. Many customers choose a $1 chocolate bar or a $1.50 doughnut over the cookie, or they simply resist the temptation of the cookie at the higher price. Before we do any math, this assumption suggests that the demand for cookies is elastic.

Adding in the numbers, we find that Helen’s weekly sales drop from 200 cookies to 150 cookies. This is a 25% change in demand on account of a 10% price increase. We immediately see that the change in demand is greater than the change in price. That means that demand is elastic. Let’s do the math.

\[
\text{percent change in quantity} = \frac{150 - 200}{(150 + 200) \div 2} \times 100 = \frac{-50}{175} \times 100 = -28.75
\]

\[
\text{percent change in price} = \frac{2.20 - 2.00}{(2.00 + 2.20) \div 2} \times 100 = \frac{0.20}{2.10} \times 100 = 9.52
\]

\[
\text{Price Elasticity of Demand} = \frac{-28.75 \text{ percent}}{9.52 \text{ percent}} = -3
\]

When the absolute value of the price elasticity is > 1, the demand is elastic. In this example, the demand for cookies is elastic.

What impact does this have on Helen’s objective to increase revenue? It’s not pretty.

- Price 1: 200 cookies sold x $2.00 per cookie = $400
- Price 2: 150 cookies sold x $2.20 = $330

She is earning less revenue because of the price change. What should Helen do next? She has learned that a small change in price leads to a large change in demand. What if she lowered the price slightly from her original $2.00 price? If the pattern holds, then a small reduction in price will lead to a large increase in sales. That would give her a much more favorable result.

---

**Try It**

Visit this page in your course online to view this presentation.

---

**Try It**

These questions allow you to get as much practice as you need, as you can click the link at the top of the first question (“Try another version of these questions”) to get a new set of questions. Practice until you feel comfortable doing the questions.

Visit this page in your course online to practice before taking the quiz.
ELASTICITY, COSTS, AND CUSTOMERS

Learning Objectives

- Evaluate how elasticity can cause shifts in demand and supply
- Predict how elasticity affects equilibrium in the long-run and the short-run

Customers and Changing Costs

We can see that understanding elasticity helps a firm set a price that maximizes total revenue. What happens if the firm’s production costs change, though? And what is the impact on customers?

Most businesses are continually trying to figure out ways to produce at a lower cost, as one path to earning higher profits. It’s a challenge to do this, though, when the price of a key input over which a firm has no control rises. For example, many chemical companies use petroleum as a key input, but they have no control over the world market price for crude oil. Coffee shops use coffee as a key input, but they have no control over the world market price of coffee. If the cost of a key input rises, can the firm pass along those higher costs to consumers in the form of higher prices?

Conversely, if new and less expensive ways of producing are invented, can the firm keep the benefits in the form of higher profits, or will the market pressure them to pass along the gains to consumers in the form of lower prices? The price elasticity of demand plays a key role in answering these questions.

Imagine that, as a consumer of legal pharmaceutical products, you read a news story about a technological breakthrough in the production of aspirin: now every aspirin factory can make aspirin more cheaply than it did before.

What does this discovery mean to you? Figure 2 illustrates two possibilities. In Figure 2(a), the demand curve is highly inelastic. In this case, a technological breakthrough that shifts supply to the right, from $S_0$ to $S_1$, so that the equilibrium shifts from $E_0$ to $E_1$, creates a substantially lower price for the product with relatively little impact on the quantity sold.

In Figure 2(b), the demand curve is highly elastic. In this case, the technological breakthrough leads to a much greater quantity sold in the market at very close to the original price. Consumers benefit more, in general, when the demand curve is more inelastic because the shift in the supply results in a much lower price for consumers.
Figure 1. Pills and Production. Consider how a technological breakthrough in aspirin production would impact the market if it were inelastic and if it were elastic.

Figure 2. Passing along Cost Savings to Consumers. Cost-saving gains cause supply to shift out to the right from $S_0$ to $S_1$; that is, at any given price, firms will be willing to supply a greater quantity. If demand is inelastic, as in (a), the result of this cost-saving technological improvement will be substantially lower prices. If demand is elastic, as in (b), the result will be only slightly lower prices. Consumers benefit in either case, from a greater quantity at a lower price, but the benefit is greater when demand is inelastic, as in (a).

Producers of aspirin may find themselves in a nasty bind here. The situation shown in Figure 2, with extremely inelastic demand, means that a new invention may cause the price to drop dramatically while quantity changes little. As a result, the new production technology can lead to a drop in the revenue that firms earn from sales of aspirin. However, if strong competition exists between producers of aspirin, each producer may have little choice but to search for and implement any breakthrough that allows it to reduce production costs.

After all, if one firm decides not to implement such a cost-saving technology, it can be driven out of business by other firms that do.
Since demand for food is generally inelastic, farmers may often face the situation in Figure 2(a). That is, a surge in production leads to a severe drop in price that can actually decrease the total revenue received by farmers. Conversely, poor weather or other conditions that cause a terrible year for farm production can sharply raise prices so that the total revenue received increases. The example below discusses how these issues relate to coffee.

**Fluctuations in Coffee Prices**

Coffee is an international crop. The top five coffee-exporting nations are Brazil, Vietnam, Colombia, Indonesia, and Guatemala. In these nations and others, 20 million families depend on selling coffee beans as their main source of income. These families are exposed to enormous risk, because the world price of coffee bounces up and down. For example, in 1993, the world price of coffee was about 50 cents per pound; in 1995 it was four times as high, at $2 per pound. By 1997 it had fallen by half to $1.00 per pound. In 1998 it leaped back up to $2 per pound. By 2001 it had fallen back to 46 cents a pound; by early 2011 it went back up to about $2.31 per pound. By the end of 2012, the price had fallen back to about $1.31 per pound.

The reason for these price fluctuations stems from a combination of inelastic demand and shifts in supply. The elasticity of coffee demand is only about 0.3; that is, a 10% rise in the price of coffee leads to a decline of about 3% in the quantity of coffee consumed. When a major frost hit the Brazilian coffee crop in 1994, coffee supply shifted to the left with an inelastic demand curve, leading to much higher prices. Conversely, when Vietnam entered the world coffee market as a major producer in the late 1990s, the supply curve shifted out to the right. With a highly inelastic demand curve, coffee prices fell dramatically. This situation is shown in Figure 2 (a), above.

**Watch It**

Get some more practice applying elasticity in the real world by watching the following video about what happens to the housing rental market (and the prices you would pay for rent) depending on whether or not demand is elastic or inelastic.

Watch this video online: [https://youtu.be/EZl86UOlOkg](https://youtu.be/EZl86UOlOkg)

Elasticity also reveals whether firms can pass higher costs that they incur on to consumers. Addictive substances, for which demand is inelastic, are products for which producers can pass higher costs on to consumers. For example, the demand for cigarettes is relatively inelastic among regular smokers who are somewhat addicted. Economic research suggests that increasing cigarette prices by 10% leads to about a 3% reduction in the quantity of cigarettes that adults smoke, so the elasticity of demand for cigarettes is 0.3. If society increases taxes on companies that produce cigarettes, the result will be, as in Figure 3(a), that the supply curve shifts from S₀ to S₁. However, as the equilibrium moves from E₀ to E₁, producers mainly pass along these taxes to consumers in the form of higher prices. These higher taxes on cigarettes will raise tax revenue for the government, but they will not much affect the quantity of smoking.

If the goal is to reduce the quantity of cigarettes demanded, we must achieve it by shifting this inelastic demand back to the left, perhaps with public programs to discourage cigarette use or to help people to quit. For example, anti-smoking advertising campaigns have shown some ability to reduce smoking. However, if cigarette demand were more elastic, as in Figure 3 (b), then an increase in taxes that shifts supply from S₀ to S₁ and equilibrium from E₀ to E₁ would reduce the quantity of cigarettes smoked substantially. Youth smoking seems to be more elastic than adult smoking—that is, the quantity of youth smoking will fall by a greater percentage than the quantity of adult smoking in response to a given percentage increase in price.
Figure 3. Passing along Higher Costs to Consumers. Higher costs, like a higher tax on cigarette companies for the example given in the text, lead supply to shift to the left. This shift is identical in (a) and (b). However, in (a), where demand is inelastic, the cost increase can largely be passed along to consumers in the form of higher prices, without much of a decline in equilibrium quantity. In (b), demand is elastic, so the shift in supply results primarily in a lower equilibrium quantity. Consumers suffer in either case, but in (a), they suffer from paying a higher price for the same quantity, while in (b), they suffer from buying a lower quantity (and presumably needing to shift their consumption elsewhere).

Try It
Visit this page in your course online to view this presentation.

Long-Run vs. Short-Run Impact

Elasticities are often lower in the short run than in the long run. On the demand side of the market, it can sometimes be difficult to change Qd in the short run but easier in the long run. Consumption of energy is a clear example. In the short run, it is not easy for a person to make substantial changes in his or her energy consumption. Maybe you can carpool to work sometimes or adjust your home thermostat by a few degrees if the cost of energy rises, but that’s about it. However, in the long-run you can purchase a car that gets more miles to the gallon, choose a job that is closer to where you live, buy more energy-efficient home appliances, or install more insulation in your home.

As a result, the elasticity of demand for energy is somewhat inelastic in the short run, but much more elastic in the long run. Figure 4 shows an example, based roughly on historical experience, of the responsiveness of Qd to price changes. In 1973, the price of crude oil was $12 per barrel, and total consumption in the U.S. economy was 17 million barrels per day. That year, the nations who were members of the Organization of Petroleum Exporting Countries (OPEC) cut off oil exports to the United States for six months because the Arab members of OPEC disagreed with the U.S. support for Israel. OPEC did not bring exports back to their earlier levels until 1975—a policy that can be interpreted as a shift of the supply curve to the left in the U.S. petroleum market. Figure 4(a) and (b) show the same original equilibrium point and the same identical shift of a supply curve to the left from S0 to S1.
Figure 4. How a Shift in Supply Can Affect Price or Quantity. The intersection ($E_D$) between demand curve $D$ and supply curve $S_0$ is the same in both (a) and (b). The shift of supply to the left from $S_0$ to $S_1$ is identical in both (a) and (b). The new equilibrium ($E_1$) has a higher price and a lower quantity than the original equilibrium ($E_0$) in both (a) and (b). However, the shape of the demand curve $D$ is different in (a) and (b). As a result, the shift in supply can result either in a new equilibrium with a much higher price and an only slightly smaller quantity, as in (a), or in a new equilibrium with only a small increase in price and a relatively larger reduction in quantity, as in (b).

Figure 4(a) shows inelastic demand for oil in the short run similar to that which existed for the United States in 1973. In Figure 4(a), the new equilibrium ($E_1$) occurs at a price of $25 per barrel, roughly double the price before the OPEC shock, and an equilibrium quantity of 16 million barrels per day. Figure 3(b) shows what the outcome would have been if the U.S. demand for oil had been more elastic, a result more likely over the long term. This alternative equilibrium ($E_1$) would have resulted in a smaller price increase to $14 per barrel and larger reduction in equilibrium quantity to 13 million barrels per day. In 1983, for example, U.S. petroleum consumption was 15.3 million barrels a day, which was lower than in 1973 or 1975. U.S. petroleum consumption was down even though the U.S. economy was about one-fourth larger in 1983 than it had been in 1973. The primary reason for the lower quantity was that higher energy prices spurred conservation efforts, and after a decade of home insulation, more fuel-efficient cars, more efficient appliances and machinery, and other fuel-conserving choices, the demand curve for energy had become more elastic.

On the supply side of markets, producers of goods and services typically find it easier to expand production in the long term of several years rather than in the short run of a few months. After all, in the short run it can be costly or difficult to build a new factory, hire many new workers, or open new stores. But over a few years, all of these are possible.

Indeed, in most markets for goods and services, prices bounce up and down more than quantities in the short run, but quantities often move more than prices in the long run. The underlying reason for this pattern is that supply and demand are often inelastic in the short run, so that shifts in either demand or supply can cause a relatively greater change in prices. But since supply and demand are more elastic in the long run, the long-run movements in prices are more muted, while quantity adjusts more easily in the long run.

Try It

Visit this page in your course online to view this presentation.
TAX INCIDENCE

Learning Objectives

- Explain how the price elasticities of demand and supply determine the incidence of a tax on buyers and sellers.

Elasticity and Tax Incidence

People often assume that when government imposes a tax on purchases of some product, producers simply raise the price of the product so that consumers end up paying the tax. Makes sense, right? Except like many economic myths, it’s not true. The analysis, or manner, of how a tax burden is divided between consumers and producers is called tax incidence. Tax incidence depends on the price elasticities of supply and demand.

The example of cigarette taxes introduced previously demonstrated that because demand is inelastic, taxes are not effective at reducing the equilibrium quantity of smoking, and they mainly pass along to consumers in the form of higher prices. With other products, however, the burden of the tax can be very different. Let’s drill down into these ideas.

Watch It

This video introduces the idea of the tax burden and demonstrates how taxes impact both consumers and producers. Look closely at the graphs towards the end of the video to graphically see how different elasticities cause the tax incidence to shift. When the demand is inelastic, consumers pay more of the tax, but when demand is elastic, the burden falls on the producers.

Watch this video online: https://youtu.be/9gwTH4Yme8I

Typically, the tax incidence, or burden, falls both on the consumers and producers of the taxed good. However, if one wants to predict which group will bear most of the burden, all one needs to do is examine the elasticity of demand and supply. In the tobacco example, the tax burden falls on the most inelastic side of the market.

Note also, that when taxes on sales affect the equilibrium quantity, there are effects on economic welfare. You can see that as reductions in consumer surplus, reductions in producer surplus and deadweight loss. The size of these changes depends on the price elasticities of demand and supply.

Let’s consider another example. Imagine a $1 tax on every barrel of apples that an apple farmer produces. If the product (apples) is price inelastic to the consumer then the farmer is able to pass the entire tax on to consumers of apples by raising the price by $1. In this situation, consumers bear the entire burden of the tax, or the tax incidence falls on consumers. On the other hand, if the apple farmer is unable to raise prices because the product is price elastic, the farmer has to bear the burden of the tax through decreased revenues, therefore the tax incidence falls on the farmer. If
the apple farmer can raise prices by an amount less than $1, then consumers and the farmer are sharing the tax burden. If demand is more inelastic than supply, consumers bear most of the tax burden, and if supply is more inelastic than demand, sellers bear most of the tax burden.

The intuition for this is simple. When the demand is inelastic, consumers are not very responsive to price changes, and the quantity demanded reduces only modestly when the tax is introduced. In the case of smoking, the demand is inelastic because consumers are addicted to the product. The government can then pass the tax burden along to consumers in the form of higher prices, without much of a decline in the equilibrium quantity.

Similarly, when a government introduces a tax in a market with an inelastic supply, such as, for example, beachfront hotels, and sellers have no alternative than to accept lower prices for their business, taxes do not greatly affect the equilibrium quantity. The tax burden now passes on to the sellers. If the supply was elastic and sellers had the possibility of reorganizing their businesses to avoid supplying the taxed good, the tax burden on the sellers would be much smaller. The tax would result in a much lower quantity sold instead of lower prices received. Figure 1 illustrates this relationship between the tax incidence and elasticity of demand and supply.

In Figure 1(a), the supply is inelastic and the demand is elastic, such as in the example of beachfront hotels. While consumers may have other vacation choices, sellers can’t easily move their businesses. By introducing a tax, the government essentially creates a wedge between the price paid by consumers \( P_c \) and the price received by producers \( P_p \). The distance between \( P_c \) and \( P_p \) is the tax rate. The new market price is \( P_c \), but sellers receive only \( P_p \) per unit sold, as they pay \( P_c - P_p \) to the government. Since we can view a tax as raising the costs of production, this could also be represented by a leftward shift of the supply curve, where the new supply curve would intercept the demand at the new quantity \( Q_t \). For simplicity, Figure 1 omits the shift in the supply curve.

The tax revenue is given by the shaded area, which we obtain by multiplying the tax per unit by the total quantity sold \( Q_t \). The tax incidence on the consumers is given by the difference between the price paid \( P_c \) and the initial equilibrium price \( P_e \). The tax incidence on the sellers is given by the difference between the initial equilibrium price \( P_e \) and the price they receive after the tax is introduced \( P_p \). In Figure 1(a), the tax burden falls disproportionately on the sellers, and a larger proportion of the tax revenue (the shaded area) is due to the resulting lower price received by the sellers than by the resulting higher prices paid by the buyers. Figure 1(b) describes the example of the tobacco excise tax where the supply is more elastic than demand. The tax incidence now falls disproportionately on consumers, as shown by the large difference between the price they pay, \( P_c \), and the initial equilibrium price, \( P_e \). Sellers receive a lower price than before the tax, but this difference is much smaller than the change in consumers’ price. From this analysis one can also predict whether a tax is likely to create a large revenue or not. The more elastic the demand curve, the more likely that consumers will reduce quantity instead of paying higher prices. The more elastic the supply curve, the more likely that sellers will reduce the quantity sold, instead of taking lower prices. In a market where both the demand and supply are very elastic, the imposition of an excise tax generates low revenue.
Some believe that excise taxes hurt mainly the specific industries they target. For example, the medical device excise tax, in effect since 2013, has been controversial for it can delay industry profitability and therefore hamper start-ups and medical innovation. However, whether the tax burden falls mostly on the medical device industry or on the patients depends simply on the elasticity of demand and supply.

Glossary

tax incidence: distribution of the tax burden between buyers and sellers

PUTTING IT TOGETHER: ELASTICITY

Netflix Pricing Revisited

We began this module discussing a price change that Netflix imposed on its customers. Now that we understand price elasticity, we can better evaluate that case. How did the 60 percent price increase end up for Netflix? It was a very bumpy two-year ride. Before the price increase, there were about 24.6 million U.S. subscribers. After the price increase, 810,000 infuriated customers canceled their Netflix subscriptions, dropping the total number of subscribers to 23.79 million. Fast-forward to June 2013, when there were 36 million streaming Netflix subscribers in the United States. This was an increase of 11.4 million subscribers since the price increase—an average per-quarter growth of about 1.6 million. This growth is less than the 2 million-per-quarter increases Netflix experienced in the fourth quarter of 2010 and the first quarter of 2011.

During the first year after the price increase, the firm's stock price (a measure of future expectations for the firm) fell from about $300 per share to just under $54. By June 2013, the stock price had rebounded to about $200 per share—still off by more than one-third from its high, but definitely improving. What happened? Obviously, Netflix understood the law of demand. Company officials reported, when they announced the price increase, that this could result in the loss of about 600,000 existing subscribers. Using the elasticity of demand formula, it is easy to see that they expected an inelastic response:
In addition, Netflix officials had expected that the price increase would have little impact on attracting new customers. Netflix anticipated adding up to 1.29 million new subscribers in the third quarter of 2011. It is true that this was slower growth than the firm had experienced over the past year—about 2 million per quarter. Why was the estimate of customers leaving so far off? During the fourteen years after Netflix was founded, there was an increase in the number of close, but not perfect, substitutes. Consumers now had choices ranging from Vudu, Amazon Prime, Hulu, and Redbox to retail stores.

Jaime Weinman reported in *Maclean’s* that Redbox kiosks are “a five-minute drive or less from 68 percent of Americans, and it seems that many people still find a five-minute drive more convenient than loading up a movie online.” It seems that, in 2012, many consumers still preferred a physical DVD disk over streaming video. What missteps did the Netflix management make? In addition to misjudging the elasticity of demand, by failing to account for close substitutes, it seems they may have also misjudged customers’ preferences and tastes (at the time being). Yet, we now see that as the population has increased, the preference for streaming video has overtaken the desire for physical DVD discs. Netflix, the target of numerous late-night talk-show jabs and laughs in 2011, may have had the last laugh in the end.
MODULE 6: UTILITY

WHY IT MATTERS: UTILITY

Why use utility to explain consumer behavior?

Figure 1. Consumption Collage. All images by Chris Potter, CC-BY. Best Buy, Costco, Macy’s, Home Depot

The 2008–2009 Great Recession touched families around the globe. In too many countries, workers found themselves out of a job. In developed countries, unemployment compensation provided a safety net, but families still saw a marked decrease in disposable income and had to make tough spending decisions. Of course, non-essential, discretionary spending was the first to go.

Even so, there was one particular category that saw a universal increase in spending world-wide during that time—an 18% uptick in the United States, specifically. You might guess that consumers began eating more meals at home, increasing grocery store spending; however, the Bureau of Labor Statistics’ Consumer Expenditure Survey, which tracks U.S. food spending over time, showed “real total food spending by U.S. households declined five percent between 2006 and 2009.” So, it was not groceries. What product would people around the world demand more of during tough economic times, and more importantly, why? (Find out at module’s end.)

That question leads us to this module’s topic—analyzing how consumers make choices.

As a student you may be interested in maximizing your grade point average (GPA) in order to qualify for scholarships, transfer to a different college or university, or even go to graduate school. In order to make your GPA as high as you can, which of the following two strategies should you choose?
• Spend equal amounts of time studying for each of your classes, or
• Study more for some classes and less for others

Generally, the answer is that you should be studying more for some classes and less for others. Why?

This module introduces the “biggest bang for the buck” principle, which is one of the key principles of microeconomics. This principle explains how to make the best choices you can in conditions of scarcity. For example, if you only have a few hours you can spend studying each day, how should you divide them up between your different classes?

![Figure 2. The Consumer’s Mind by Christopher Dombres, CC-BY.](image)

This module will also help you answer questions like:

• Why do people purchase more of something when its price falls?
• Why do people buy more goods and services when their budget increases?

In the first portion of this module, we will assume that consumers are economically rational – this is a key assumption of mainstream microeconomics. Toward the end of the module we relax this assumption as we consider an alternative approach called behavioral economics. Behavioral economics acknowledges that sometimes we don’t make the best choices.

**INTRODUCTION TO UTILITY AND CONSUMER EQUILIBRIUM**

What you’ll learn to do: describe the concept of utility and explain how consumers spend in order to maximize utility
Microeconomics seeks to understand the behavior of individual economic agents such as individuals and businesses. Economists believe that we can analyze individuals’ decisions, such as what goods and services to buy, as choices we make within certain budget constraints. Generally, consumers are trying to get the most for their limited budget. In economic terms they are trying to maximize total utility, or satisfaction, given their budget constraint.

Everyone has their own personal tastes and preferences. The French say: *Chacun à son goût*, or “Each to his own taste.” An old Latin saying states, *De gustibus non est disputandum* or “There’s no disputing about taste.” If people base their decisions on their own tastes and personal preferences, however, then how can economists hope to analyze the choices consumers make?

An economic explanation for why people make different choices begins with accepting the proverbial wisdom that tastes are a matter of personal preference. However, economists also believe that the choices people make are influenced by their incomes, by the prices of goods and services they consume, and by factors like where they live.

This section introduces the economic theory of how consumers make choices about what goods and services to buy with their limited income. If you look at Facebook for five fewer minutes a day, will you really be more productive? Is it worth it to clock into work ten minutes early, or would it be best to spend that time with your significant other? Is it worth it to spend five dollars on a dessert when you already feel a little bit full? In this section, you will examine choices made at the margin, or the decisions you make to do a little more or a little less of something.
Consumer Choices

Information on the consumption choices of Americans is available from the Consumer Expenditure Survey carried out by the U.S. Bureau of Labor Statistics. Figure 1 shows spending patterns for the average U.S. household. The first row shows income and, after taxes and personal savings are subtracted, it shows that, in 2015, the average U.S. household spent $48,109 on consumption. The table then breaks down consumption into various categories. The average U.S. household spent roughly one-third of its consumption on shelter and other housing expenses, another one-third on food and vehicle expenses, and the rest on a variety of items, as shown. These patterns will vary for specific households by differing levels of family income, by geography, and by preferences.

<table>
<thead>
<tr>
<th>Average Household Income before Taxes</th>
<th>$62,481</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Annual Expenditures</td>
<td>$48,109</td>
</tr>
<tr>
<td>Food at home</td>
<td>$3,264</td>
</tr>
<tr>
<td>Food away from home</td>
<td>$2,505</td>
</tr>
<tr>
<td>Housing</td>
<td>$16,557</td>
</tr>
<tr>
<td>Apparel and services</td>
<td>$1,700</td>
</tr>
<tr>
<td>Transportation</td>
<td>$7,677</td>
</tr>
<tr>
<td>Healthcare</td>
<td>$3,157</td>
</tr>
<tr>
<td>Entertainment</td>
<td>$2,504</td>
</tr>
<tr>
<td>Education</td>
<td>$1,074</td>
</tr>
<tr>
<td>Personal insurance and pensions</td>
<td>$5,357</td>
</tr>
<tr>
<td>All else: alcohol, tobacco, reading, personal care, cash contributions, miscellaneous</td>
<td>$3,356</td>
</tr>
</tbody>
</table>

Consumer Choice and Utility

When economists talk about consumer choice, what they are referring to is the combination of goods and services a consumer purchases. To understand how a household will make its choices, economists look at what consumers can afford, as shown in a budget constraint (or budget line), and the total utility or satisfaction derived from those choices. When we graph a budget constraint, the quantity of one good is on the horizontal axis and the quantity of the other good on the vertical axis. The budget constraint line shows the various combinations of two goods that are affordable given a specific budget (or level of consumer income).

Watch It

Watch the selected clip from this video to review budget constraints. The budget constraint graph shows the various combinations of goods (like coffee and pizza) that you can afford at any given price. We'll finish the second portion of the video later in the module.

Visit this page in your course online to view this presentation.
Consider José’s situation. José likes to collect T-shirts and watch movies. In Figure 1 we show the quantity of T-shirts on the horizontal axis while we show the quantity of movies on the vertical axis. If José had unlimited income or goods were free, then he could consume without limit. However, José, like all of us, faces a budget constraint. José has a total of $56 to spend. The price of T-shirts is $14 and the price of movies is $7. Notice that the vertical intercept of the budget constraint line is at eight movies and zero T-shirts ($56/7=8). The horizontal intercept of the budget constraint is four, where José spends all of his money on T-shirts and no movies ($56/14=4). The slope of the budget constraint line is rise/run or $-8/4=-2$. The specific choices along the budget constraint line show the combinations of affordable T-shirts and movies.

![Figure 1. A Choice between Consumption Goods. José has income of $56. Movies cost $7 and T-shirts cost $14. The points on the budget constraint line show the combinations of affordable movies and T-shirts.](image)

**Utility** is the term economists use to describe the satisfaction or happiness a person gets from consuming a good or service. José obtains utility from consuming T-shirts and consuming movies. Like all consumers, we assume José wishes to choose the combination of T-shirts and movies that will provide him with the greatest total utility.

Let's begin with an assumption, which we will discuss in more detail later, that José can measure his own utility with something called *utils*. (It is important to note that you cannot make comparisons between the utils of individuals. If one person gets 20 utils from a cup of coffee and another gets 10 utils, this does not mean than the first person gets more enjoyment from the coffee than the other or that they enjoy the coffee twice as much. The reason why is that utils are subjective to an individual. The way one person measures utils is not the same as the way someone else does.)

<table>
<thead>
<tr>
<th>T-Shirts (Quantity)</th>
<th>Total Utility</th>
<th>Movies (Quantity)</th>
<th>Total Utility</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>22</td>
<td>1</td>
<td>16</td>
</tr>
<tr>
<td>2</td>
<td>43</td>
<td>2</td>
<td>31</td>
</tr>
<tr>
<td>3</td>
<td>63</td>
<td>3</td>
<td>45</td>
</tr>
<tr>
<td>4</td>
<td>81</td>
<td>4</td>
<td>58</td>
</tr>
<tr>
<td>5</td>
<td>97</td>
<td>5</td>
<td>70</td>
</tr>
<tr>
<td>6</td>
<td>111</td>
<td>6</td>
<td>81</td>
</tr>
<tr>
<td>7</td>
<td>123</td>
<td>7</td>
<td>91</td>
</tr>
<tr>
<td>8</td>
<td>133</td>
<td>8</td>
<td>100</td>
</tr>
</tbody>
</table>

Table 2 shows how José’s utility is connected with his T-shirt or movie consumption. The first column of the table shows the quantity of T-shirts consumed. The second column shows the total utility, or total amount of satisfaction, that José
receives from consuming that number of T-shirts. The typical pattern of total utility, shown in this example, is that consuming additional units of a good leads to greater total utility, but at a decreasing rate. We can see this in Figure 2.

**Figure 2. Total Utility Curve.** This is a typical total utility curve showing an increase in total utility as consumption of a good increases, though at a decreasing rate.

The rest of Table 2 shows the utility José would obtain from attending different quantities of movies. Total utility follows the expected pattern: it increases as the number of movies that José watches rises. José can afford any combination of T-shirts and movies which is on his budget constraint. Which combination should he choose if he wishes to obtain the most utility possible? Table 3 looks at each point on the budget constraint in Figure 1, and adds up José’s total utility for each combination of T-shirts and movies.

<table>
<thead>
<tr>
<th>Point</th>
<th>T-Shirts</th>
<th>Movies</th>
<th>Total Utility</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>4</td>
<td>0</td>
<td>81 + 0 = 81</td>
</tr>
<tr>
<td>Q</td>
<td>3</td>
<td>2</td>
<td>63 + 31 = 94</td>
</tr>
<tr>
<td>R</td>
<td>2</td>
<td>4</td>
<td>43 + 58 = 101</td>
</tr>
<tr>
<td>S</td>
<td>1</td>
<td>6</td>
<td>22 + 81 = 103</td>
</tr>
<tr>
<td>T</td>
<td>0</td>
<td>8</td>
<td>0 + 100 = 100</td>
</tr>
</tbody>
</table>

**Calculating Total Utility**

Let's look at how José makes his decision in more detail.

**Step 1.** Observe that, at point Q in Figure 1 (for example), José consumes three T-shirts and two movies.

**Step 2.** Look at Table 2. You can see from the fourth row/second column that three T-shirts are worth 63 utils. Similarly, the second row/fifth column shows that two movies are worth 31 utils.

**Step 3.** From this information, you can calculate that point Q has a total utility of 94 (63 + 31).

**Step 4.** You can repeat the same calculations for each point on Table 3, in which the total utility numbers are shown in the last column.

For José, the highest total utility for all possible combinations of goods occurs at point S, with a total utility of 103 from consuming one T-shirt and six movies.
MARGINAL UTILITY VERSUS TOTAL UTILITY

Learning Objectives

- Differentiate between total and marginal utility
- Contrast and compute marginal utility and total utility

Choices Are Made at the Margin

Economists argue that most choices are made “at the margin.” The margin is the current level of an activity. Think of it as the edge from which a choice is to be made. A choice at the margin is a decision to do a little more or a little less of something.

Assessing choices at the margin can lead to extremely useful insights. Consider, for example, the problem of curtailing water consumption when the amount of water available falls short of the amount people now use. If you live east of the
Mississippi River in the U.S., this is not likely to be a problem for you, but it's very common in the West. (For a historical example, check out the movie *Chinatown.*) Economists argue that one way to induce people to conserve water is to raise its price. A common response to this recommendation is that a higher price would have no effect on water consumption, because water is a necessity. Many people claim that prices do not affect water consumption because people “need” water.

But choices in water consumption, like virtually all choices, are made at the margin. Individuals do not make choices about whether they should or should not consume water. Rather, they decide whether to consume a little more or a little less water. Household water consumption in the United States totals about 105 gallons per person per day. Think of that starting point as the edge from which a choice at the margin in water consumption is made. Could a higher price cause you to use less water brushing your teeth, take shorter showers, or water your lawn less? Could a higher price cause people to reduce their use, say, to 104 gallons per person per day? To 103? When we examine the choice to consume water at the margin, the notion that a higher price would reduce consumption seems much more plausible. Prices affect our consumption of water because choices in water consumption, like other choices, are made at the margin.

In exploring consumer choices, it's important to differentiate between total utility and marginal utility. Marginal utility is based on the notion that individuals rarely face all-or-nothing decisions, that most of the time they are considering a little more or a little less of something when allocating their budget, time or other scarce resources. So a student might ask, “how much better will I do on an exam if I study for one more hour?” The answer could be called the marginal grade improvement. Marginal utility, then, is the change in total utility from consuming one more or one less of an item. For example, the marginal utility of a third slice of pizza is the change in satisfaction one gets when eating the third slice instead of stopping with two.

As we continue to study microeconomics, we will see a number of additional marginal concepts. The general formula for computing a marginal outcome is the change in the outcome divided by the change in the number of inputs used to produce that outcome. For example, if two more hours of work yields an additional $20 in wages, the marginal wage earned is $20/2 hours = $10 per hour. The marginal cost of one more unit of output a firm produces is the amount that total cost increases when the firm produces one more unit of output.

Watch It

Watch this clip to understand how consumers make choices based on the utility, or satisfaction, derived from the purchase. The utility of a good or purchase decreases as a person consumes more of that good.

Watch this video online: [https://youtu.be/oG_R1jQPqTc](https://youtu.be/oG_R1jQPqTc)

### Marginal Utility vs. Total Utility

In Table 1 below, we repeat the information on total utility shown in Table 2 on the previous page, but we also include marginal utility.

<table>
<thead>
<tr>
<th>T-Shirts (Quantity)</th>
<th>Total Utility</th>
<th>Marginal Utility</th>
<th>Movies (Quantity)</th>
<th>Total Utility</th>
<th>Marginal Utility</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>22</td>
<td>22</td>
<td>1</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>2</td>
<td>43</td>
<td>21</td>
<td>2</td>
<td>31</td>
<td>15</td>
</tr>
<tr>
<td>3</td>
<td>63</td>
<td>20</td>
<td>3</td>
<td>45</td>
<td>14</td>
</tr>
<tr>
<td>4</td>
<td>81</td>
<td>18</td>
<td>4</td>
<td>58</td>
<td>13</td>
</tr>
<tr>
<td>5</td>
<td>97</td>
<td>16</td>
<td>5</td>
<td>70</td>
<td>12</td>
</tr>
<tr>
<td>6</td>
<td>111</td>
<td>14</td>
<td>6</td>
<td>81</td>
<td>11</td>
</tr>
<tr>
<td>7</td>
<td>123</td>
<td>12</td>
<td>7</td>
<td>91</td>
<td>10</td>
</tr>
<tr>
<td>8</td>
<td>133</td>
<td>10</td>
<td>8</td>
<td>100</td>
<td>9</td>
</tr>
</tbody>
</table>
The third column of Table 1 shows marginal utility, which is the additional utility provided by one additional unit of consumption. This equation for marginal utility is:

\[
MU = \frac{\text{change in total utility}}{\text{change in quantity}}
\]

Figure 1 shows the graph of marginal utility. Notice that marginal utility diminishes as additional units are consumed, which means that each subsequent unit of a good consumed provides less additional utility. For example, the first T-shirt José picks is his favorite and it gives him an addition of 22 utils. The fourth T-shirt is just something to wear when all his other clothes are in the wash and yields only 18 additional utils. This is an example of the law of diminishing marginal utility, which holds that the additional utility decreases with each unit added. Diminishing marginal utility is another example of the more general law of diminishing returns we learned earlier in the module about choices and budget constraints.

Marginal utility for movies (column 6) also follows the expected pattern: each additional movie brings a smaller gain in utility than the previous one. The first movie José attends is the one he wanted to see the most, and thus provides him with the highest level of utility or satisfaction. The fifth movie he attends is just to kill time. Notice that total utility is also the sum of the marginal utilities.

Try It

Visit this page in your course online to check your understanding.

Try It

These questions allow you to get as much practice as you need, as you can click the link at the top of the first question (“Try another version of these questions”) to get a new set of questions. Practice until you feel comfortable doing the questions.

Visit this page in your course online to practice before taking the quiz.

Glossary

**diminishing marginal utility**: the common pattern that each marginal unit of a good consumed provides less of an addition to utility than the previous unit

**marginal utility**: the additional utility provided by one additional unit of consumption
RULES FOR MAXIMIZING UTILITY

Learning Objectives

- Explain why maximizing utility requires that the last unit of each item purchased must have the same marginal utility per dollar
- Calculate the utility-maximizing choice

The problem of finding consumer equilibrium, that is, the combination of goods and services that will maximize an individual’s total utility, comes down to comparing the trade-offs between one affordable combination (shown by a point on the budget line in Figure 1, below) with all the other affordable combinations.

Most people approach their utility-maximizing combination of choices in a step-by-step way. This step-by-step approach is based on looking at the tradeoffs, measured in terms of marginal utility, of consuming less of one good and more of another. You can think of this step-by-step approach as the “biggest bang for the buck” principle. For example, say that José starts off thinking about spending all his money on T-shirts and choosing point P, which corresponds to four T-shirts and no movies, as illustrated in Figure 1.

José chooses this starting point randomly; he has to start somewhere. Then he considers giving up the last T-shirt, the one that provides him the least marginal utility, and using the money he saves to buy two movies instead. Table 1 tracks the step-by-step series of decisions José needs to make (Key: T-shirts cost $14, movies cost $7, and Jose’s income is $56).
Table 1. A Step-by-Step Approach to Maximizing Utility

<table>
<thead>
<tr>
<th>Try</th>
<th>Which Has</th>
<th>Total Utility</th>
<th>Marginal Gain and Loss of Utility, Compared with Previous Choice</th>
<th>Conclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Choice 1: P</td>
<td>4 T-shirts and 0 movies</td>
<td>81 from 4 T-shirts + 0 from 0 movies = 81</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Choice 2: Q</td>
<td>3 T-shirts and 2 movies</td>
<td>63 from 3 T-shirts + 31 from 0 movies = 94</td>
<td>Loss of 18 from 1 less T-shirt, but gain of 31 from 2 more movies, for a net utility gain of 13</td>
<td>Q is preferred over P</td>
</tr>
<tr>
<td>Choice 3: R</td>
<td>2 T-shirts and 4 movies</td>
<td>43 from 2 T-shirts + 58 from 4 movies = 101</td>
<td>Loss of 20 from 1 less T-shirt, but gain of 27 from two more movies for a net utility gain of 7</td>
<td>R is preferred over Q</td>
</tr>
<tr>
<td>Choice 4: S</td>
<td>1 T-shirt and 6 movies</td>
<td>22 from 1 T-shirt + 81 from 6 movies = 103</td>
<td>Loss of 21 from 1 less T-shirt, but gain of 23 from two more movies for a net utility gain of 2</td>
<td>S is preferred over R</td>
</tr>
<tr>
<td>Choice 5: T</td>
<td>0 T-shirts and 8 movies</td>
<td>0 from 0 T-shirts + 100 from 8 movies = 100</td>
<td>Loss of 22 from 1 less T-shirt, but gain of 19 from two more movies, for a net utility loss of 3</td>
<td>S is preferred over T</td>
</tr>
</tbody>
</table>

DECISION MAKING BY COMPARING MARGINAL UTILITY

José could use the following thought process (if he thought in utils) to make his decision regarding how many T-shirts and movies to purchase:

**Step 1.** From Table 1, José can see that the marginal utility of the fourth T-shirt is 18. If José gives up the fourth T-shirt, then he loses 18 utils.

**Step 2.** Giving up the fourth T-shirt, however, frees up $14 (the price of a T-shirt), allowing José to buy the first two movies (at $7 each).

**Step 3.** José knows that the marginal utility of the first movie is 16 and the marginal utility of the second movie is 15. Thus, if José moves from point P to point Q, he gives up 18 utils (from the T-shirt), but gains 31 utils (from the movies).

**Step 4.** Gaining 31 utils and losing 18 utils is a net gain of 13. This is just another way of saying that the total utility at Q (94 according to the last column in Table 1) is 13 more than the total utility at P (81).

**Step 5.** So, for José, it makes sense to give up the fourth T-shirt in order to buy two movies.

José clearly prefers point Q to point P. Now repeat this step-by-step process of decision making with marginal utilities. José thinks about giving up the third T-shirt and surrendering a marginal utility of 20, in exchange for purchasing two more movies that promise a combined marginal utility of 27. José prefers point R to point Q. What if José thinks about going beyond R to point S? Giving up the second T-shirt means a marginal utility loss of 21, and the marginal utility gain from the fifth and sixth movies would combine to make a marginal utility gain of 23, so José prefers point S to R.

However, if José seeks to go beyond point S to point T, he finds that the loss of marginal utility from giving up the first T-shirt is 22, while the marginal utility gain from the last two movies is only a total of 19. If José were to choose point T, his utility would fall to 100. Through these stages of thinking about marginal tradeoffs, José again concludes that S, with one T-shirt and six movies, is the choice that will provide him with the highest level of total utility. This step-by-step approach will reach the same conclusion regardless of José’s starting point.

This approach to finding consumer equilibrium is somewhat tedious. Next, we’ll turn to a quicker and more intuitive approach.

A Rule for Maximizing Utility
This process of decision making described previously suggests a rule to follow when maximizing utility. Since the price of T-shirts is not the same as the price of movies, it’s not enough to just compare the marginal utility of T-shirts with the marginal utility of movies. Instead, we need to control for the prices of each product. We can do this by computing and comparing marginal utility per dollar of expenditure for each product. **Marginal utility per dollar** is the amount of additional utility José receives given the price of the product.

\[
\text{Marginal utility per dollar} = \frac{\text{marginal utility}}{\text{price}}
\]

For José’s T-shirts and movies, the marginal utility per dollar is shown in Table 2.

<table>
<thead>
<tr>
<th>Quantity of T-Shirts</th>
<th>Total Utility of T-Shirts</th>
<th>Marginal Utility of T-Shirts</th>
<th>Marginal Utility per Dollar</th>
<th>Quantity of Movies</th>
<th>Total Utility of Movies</th>
<th>Marginal Utility of Movies</th>
<th>Marginal Utility per Dollar</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>22</td>
<td>22</td>
<td>22/14 = 1.6</td>
<td>1</td>
<td>16</td>
<td>16</td>
<td>16/7 = 2.3</td>
</tr>
<tr>
<td>2</td>
<td>43</td>
<td>21</td>
<td>21/14 = 1.5</td>
<td>2</td>
<td>31</td>
<td>15</td>
<td>15/7 = 2.14</td>
</tr>
<tr>
<td>3</td>
<td>63</td>
<td>20</td>
<td>20/14 = 1.4</td>
<td>3</td>
<td>45</td>
<td>14</td>
<td>14/7 = 2</td>
</tr>
<tr>
<td>4</td>
<td>81</td>
<td>18</td>
<td>18/14 = 1.3</td>
<td>4</td>
<td>58</td>
<td>13</td>
<td>13/7 = 1.9</td>
</tr>
<tr>
<td>5</td>
<td>97</td>
<td>16</td>
<td>16/14 = 1.1</td>
<td>5</td>
<td>70</td>
<td>12</td>
<td>12/7 = 1.7</td>
</tr>
<tr>
<td>6</td>
<td>111</td>
<td>14</td>
<td>14/14 = 1</td>
<td>6</td>
<td>81</td>
<td>11</td>
<td>11/7 = 1.6</td>
</tr>
<tr>
<td>7</td>
<td>123</td>
<td>12</td>
<td>12/14 = 1.2</td>
<td>7</td>
<td>91</td>
<td>10</td>
<td>10/7 = 1.4</td>
</tr>
</tbody>
</table>

### A Rule for maximizing Utility

If a consumer wants to maximize total utility, for every dollar that they spend, they should spend it on the item which yields the greatest marginal utility per dollar of expenditure.

### Applying the Rule

José’s first purchase will be a movie. Why? José’s choices are to purchase either a T-shirt or a movie. Table 1 shows that the marginal utility per dollar spent on the first T-shirt is 1.6 compared with 2.5 for the first movie. Because the first movie gives José more marginal utility per dollar than the first T-shirt, and because the movie is within his budget, he will purchase a movie first.

José will continue to purchase the good which gives him the highest marginal utility per dollar until he exhausts the budget. José will keep purchasing movies because they give him a greater “bang for the buck” until the sixth movie is equivalent to a T-shirt purchase. José can afford to purchase that T-shirt. So José will choose to purchase six movies and one T-shirt. That combination, six movies and one T-shirt, is his consumer equilibrium.

### Another Rule for Maximizing Utility

Since the price of T-shirts is twice as high as the price of movies, to maximize utility the last T-shirt chosen needs to provide exactly twice the marginal utility (MU) of the last movie. If the last T-shirt provides less than twice the marginal utility of the last movie, then the T-shirt is providing less “bang for the buck” (i.e., marginal utility per dollar spent) than if the same money were spent on movies. If this is so, José should trade the T-shirt for more movies to increase his total utility. Marginal utility per dollar measures the additional utility that José will enjoy given what he has to pay for the good.

If the last T-shirt provides more than twice the marginal utility of the last movie, then the T-shirt is providing more “bang for the buck” or marginal utility per dollar, than if the money were spent on movies. As a result, José should buy more T-
shirts. Notice that at José’s optimal choice the marginal utility from the first T-shirt, of 22 is exactly twice the marginal utility of the sixth movie, which is 11. At this choice, the marginal utility per dollar is the same for both goods. This is a tell-tale signal that José has found the point with highest total utility.

This argument can be written as another rule: the utility-maximizing choice between consumption goods occurs where the marginal utility per dollar is the same for both goods, and the consumer has exhausted his or her budget.

\[ \frac{MU_1}{P_1} = \frac{MU_2}{P_2} \]

A sensible economizer will pay twice as much for something only if, in the marginal comparison, the item confers twice as much utility. Notice that the formula for the table above is

\[ \frac{22}{14} = \frac{11}{7} \]

\[ 1.6 = 1.6 \]

The following feature provides step-by-step guidance for this concept of utility-maximizing choices.

### Another Rule for Maximizing Utility

The rule, \( \frac{MU_1}{P_1} = \frac{MU_2}{P_2} \), means that the last dollar spent on each good provides exactly the same marginal utility. So:

**Step 1.** If we traded a dollar more of movies for a dollar more of T-shirts, the marginal utility gained from T-shirts would exactly offset the marginal utility lost from fewer movies. In other words, the net gain would be zero.

**Step 2.** Products, however, usually cost more than a dollar, so we cannot trade a dollar’s worth of movies. The best we can do is trade two movies for another T-shirt, since in this example T-shirts cost twice what a movie does.

**Step 3.** If we trade two movies for one T-shirt, we would end up at point R (two T-shirts and four movies).

**Step 4.** Choice 4 in Table 3 shows that if we move to point S, we would lose 21 utils from one less T-shirt, but gain 23 utils from two more movies, so we would end up with more total utility at point S.

<table>
<thead>
<tr>
<th>Try</th>
<th>Which Has</th>
<th>Total Utility</th>
<th>Marginal Gain and Loss of Utility, Compared with Previous Choice</th>
<th>Conclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Choice 1: P</td>
<td>4 T-shirts and 0 movies</td>
<td>81 from 4 T-shirts + 0 from 0 movies = 81</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Choice 2: Q</td>
<td>3 T-shirts and 2 movies</td>
<td>63 from 3 T-shirts + 31 from 0 movies = 94</td>
<td>Loss of 18 from 1 less T-shirt, but gain of 31 from 2 more movies, for a net utility gain of 13</td>
<td>Q is preferred over P</td>
</tr>
<tr>
<td>Choice 3: R</td>
<td>2 T-shirts and 4 movies</td>
<td>43 from 2 T-shirts + 58 from 4 movies = 101</td>
<td>Loss of 20 from 1 less T-shirt, but gain of 27 from two more movies for a net utility gain of 7</td>
<td>R is preferred over Q</td>
</tr>
<tr>
<td>Choice 4: S</td>
<td>1 T-shirt and 6 movies</td>
<td>22 from 1 T-shirt + 81 from 6 movies = 103</td>
<td>Loss of 21 from 1 less T-shirt, but gain of 23 from two more movies, for a net utility gain of 2</td>
<td>S is preferred over R</td>
</tr>
<tr>
<td>Choice 5: T</td>
<td>0 T-shirts and 8 movies</td>
<td>0 from 0 T-shirts + 100 from 8 movies = 100</td>
<td>Loss of 22 from 1 less T-shirt, but gain of 19 from two more movies, for a net utility loss of 3</td>
<td>S is preferred over T</td>
</tr>
</tbody>
</table>

In short, the rule shows us the utility-maximizing choice.
There is another, equivalent way to think about this. The rule can also be expressed as the ratio of the prices of the two goods should be equal to the ratio of the marginal utilities. When the price of good 1 is divided by the price of good 2, at the utility-maximizing point this will equal the marginal utility of good 1 divided by the marginal utility of good 2. This rule can be written in algebraic form:

\[
\frac{P_1}{P_2} = \frac{MU_1}{MU_2}
\]

Along the budget constraint, the total price of the two goods remains the same, so the ratio of the prices does not change. However, the marginal utility of the two goods changes with the quantities consumed. At the optimal choice of one T-shirt and six movies, point S, the ratio of marginal utility to price for T-shirts (22:14) matches the ratio of marginal utility to price for movies (of 11:7).

**Watch It**

This video applies these same concepts to a graph. It demonstrates how there will come a point when the marginal utility per dollar of one good goes down enough so that it makes more sense to buy another good.

Watch this video online: [https://youtu.be/KbW6OiuRa1Y](https://youtu.be/KbW6OiuRa1Y)

**Measuring Utility with Numbers**

This discussion of utility started off with an assumption that it is possible to place numerical values on utility, an assumption that may seem questionable. You can buy a thermometer for measuring temperature at the hardware store, but what store sells an “utilimometer” for measuring utility? However, while measuring utility with numbers is a convenient assumption to clarify the explanation, the key assumption is not that utility can be measured by an outside party, but only that individuals can decide which of two alternatives they prefer.

To understand this point, think back to the step-by-step process of finding the choice with highest total utility by comparing the marginal utility that is gained and lost from different choices along the budget constraint. As José compares each choice along his budget constraint to the previous choice, what matters is not the specific numbers that he places on his utility—or whether he uses any numbers at all—but only that he personally can identify which choices he prefers.

In this way, the step-by-step process of choosing the highest level of utility resembles rather closely how many people make consumption decisions. We think about what will make us the happiest; we think about what things cost; we think about buying a little more of one item and giving up a little of something else; we choose what provides us with the greatest level of satisfaction. The vocabulary of comparing the points along a budget constraint and total and marginal utility is just a set of tools for discussing this everyday process in a clear and specific manner. It is welcome news that specific utility numbers are not central to the argument, since a good utilimometer is hard to find. Do not worry—while we cannot measure utils, by the end of the next module, we will have transformed our analysis into something we can measure—demand.

**Try It**

Visit this page in your course online to check your understanding.

**Try It**

These questions allow you to get as much practice as you need, as you can click the link at the top of the first question (“Try another version of these questions”) to get a new set of questions. Practice until you feel comfortable doing the questions.

Visit this page in your course online to practice before taking the quiz.
Try It

These questions allow you to get as much practice as you need, as you can click the link at the top of the first question (“Try another version of these questions”) to get a new set of questions. Practice until you feel comfortable doing the questions.
Visit this page in your course online to practice before taking the quiz.

Glossary

**consumer equilibrium**: the combination of goods and services that will maximize an individual’s total utility

**marginal utility per dollar**: the additional satisfaction gained from purchasing a good given the price of the product; $\frac{MU}{Price}$

LEARN BY DOING: MAXIMIZING UTILITY

Try It

Click through the buttons in this simulation to think about applications of utility.
First, you’ll imagine that you are desperately searching for food and water on a hot day. Given the heat and your time constraints, you’ll need to decide what to obtain first based on its utility. Then, you can see these combinations on a graph and consider what the utility of each choice would be.
Be sure to choose the option to “learn more about marginal utility per dollar spent” so that you can see how marginal-decision making comes into play when choosing to forage for either apples or berries.
Visit this page in your course online to use this simulation.

INTRODUCTION TO CHANGES IN CONSUMER EQUILIBRIUM
Imagine you were to create a pie chart of your college budget and spending. Where does most of your money go? Is it spent on housing? Tuition? Maybe on a car payment or food? How much money goes towards savings, entertainment, or a night out with friends? In this section, you’ll see how consumer choices change when one’s budget changes or the prices of the things one buys changes.

**INCOME CHANGES AND CONSUMPTION CHOICES**

**Learning Objectives**

- Demonstrate how changes in income and price affect consumer choices
- Contrast the substitution effect and the income effect

Consumer equilibrium, that is, the combination of goods and services that will maximize a consumer’s utility, depends on the consumers tastes and preferences, as expressed by his or her marginal utility numbers, the prices of those items and and the budget (or income) the consumer has. In this section, we will explore how changes in a consumer’s income and changes in the prices of goods and services affect consumer choice. Then, because the budget constraint framework can be used to analyze how quantities demanded change because of price movements, the budget constraint model can illustrate the underlying logic behind demand curves.

**How Changes in Income Affect Consumer Choices**

Let’s begin with a concrete example illustrating how changes in income level affect consumer choices. Figure 1 shows a budget constraint that represents Jazmin’s choice between concert tickets at $50 each and getting away overnight to a bed-and-breakfast for $200 per night. Jazmin has $1,000 per year to spend between these two choices. Jazmin could spend all of her budget on concert tickets, in which case she could buy $1000/$50 = 20 concert tickets. This shows the vertical axis of the budget constraint (zero B&B nights and 20 concert tickets). Alternatively, she could spend all of her
budget on nights at a bed-and-breakfast, in which case she could afford $1000/200 = 5 nights. This shows the horizontal axis of her budget constraint (5 B&B nights and zero concert tickets.

After thinking about her total utility and marginal utility and applying the decision rules for maximizing utility explained earlier, Jazmin chooses point M, with eight concerts and three overnight getaways as her utility-maximizing choice.

![Figure 1. How a Change in Income Affects Consumption Choices.](image)

The utility-maximizing choice on the original budget constraint is M. The dashed horizontal and vertical lines extending through point M allow you to see at a glance whether the quantity consumed of goods on the new budget constraint is higher or lower than on the original budget constraint. On the new budget constraint, a choice like N will be made if both goods are normal goods. If overnight stays is an inferior good, a choice like P will be made. If concert tickets are an inferior good, a choice like Q will be made.

Now, assume that the income Jazmin has to spend on these two items rises to $2,000 per year. Since the prices of the two products haven't changed, doubling her budget allows Jazmin to purchase twice as many of each. This can be shown as a rightward shift in the budget constraint with intercepts as (0, 40) and (10, 0). How does this rise in income alter her utility-maximizing choice? Jazmin will again consider the total utility and marginal utility that she receives from concert tickets and overnight getaways to identify the consumer equilibrium on the new budget line. But how will her new choice relate to her original choice?

The possible choices along the new budget constraint can be divided into three groups, which are divided up by the dashed horizontal and vertical lines that pass through the original choice M in the figure. All choices on the upper left of the new budget constraint that are to the left of the vertical dashed line, like choice P with two overnight stays and 32 concert tickets, involve less of the good on the horizontal axis but much more of the good on the vertical axis. All choices to the right of the vertical dashed line and above the horizontal dashed line—like choice N with five overnight getaways and 20 concert tickets—have more consumption of both goods. Finally, all choices that are to the right of the vertical dashed line but below the horizontal dashed line, like choice Q with four concerts and nine overnight getaways, involve less of the good on the vertical axis but much more of the good on the horizontal axis.

All of these choices are theoretically possible, depending on Jazmin's personal preferences as expressed through the total and marginal utility she would receive from consuming these two goods. When income rises, the most common reaction is to purchase more of both goods, like choice N, which is to the upper right relative to Jazmin's original choice M, although exactly how much more of each good will vary according to personal taste. Conversely, when income falls, the most typical reaction is to purchase less of both goods. As we learned already in the module on elasticity, goods and services are called **normal goods** if a rise in income leads to a rise in the quantity consumed of that good and a fall in income leads to a fall in quantity consumed.

However, depending on Jazmin's preferences, a rise in income could cause consumption of one good to increase while consumption of the other good declines. A choice like P means that a rise in income caused her quantity consumed of overnight stays to decline, while a choice like Q would mean that a rise in income caused her quantity of concerts to...
Goods where demand declines as income rises (or conversely, where the demand rises as income falls) are called “inferior goods.” An inferior good occurs when people trim back on a good as income rises, because they can now afford the more expensive choices that they prefer. For example, a higher-income household might eat fewer hamburgers or be less likely to buy a used car, and instead eat more steak and buy a new car.

![Figure 2. How a Change in Price Affects Consumption Choices.](image)

**Figure 2. How a Change in Price Affects Consumption Choices.** The original utility-maximizing choice is M. When the price rises, the budget constraint shifts in to the left. The dashed lines make it possible to see at a glance whether the new consumption choice involves less of both goods, or less of one good and more of the other. The new possible choices would be fewer baseball bats and more cameras, like point H, or less of both goods, as at point J. Choice K would mean that the higher price of bats led to exactly the same quantity of bats being consumed, but fewer cameras. Choices like L are ruled out as theoretically possible but highly unlikely in the real world, because they would mean that a higher price for baseball bats means a greater quantity consumed of baseball bats.

After the price increase, Sergei will make a choice along the new budget constraint. Again, his choices can be divided into three segments by the dashed vertical and horizontal lines. In the upper left portion of the new budget constraint, at a choice like H, Sergei consumes more cameras and fewer bats. In the central portion of the new budget constraint, at a choice like J, he consumes less of both goods. At the right-hand end, at a choice like L, he consumes more bats but fewer cameras.

The typical response to higher prices is that a person chooses to consume less of the product with the higher price. This occurs for two reasons, and both effects can occur simultaneously. The **substitution effect** occurs when a price changes and consumers have an incentive to consume less of the good with a relatively higher price and more of the good with a relatively lower price. The **income effect** is that a higher price means, in effect, the buying power of income
has been reduced (even though actual income has not changed), which leads to buying less of the good (when the good is normal). In this example, the higher price for baseball bats would cause Sergei to buy fewer bats for both reasons. Exactly how much will a higher price for bats cause Sergei consumption of bats to fall? Figure 2 suggests a range of possibilities. Sergei might react to a higher price for baseball bats by purchasing the same quantity of bats, but cutting his consumption of cameras. This choice is the point K on the new budget constraint, straight below the original choice M. Alternatively, Sergei might react by dramatically reducing his purchases of bats and instead buy more cameras.

The key is that it would be imprudent to assume that a change in the price of baseball bats will only or primarily affect the good whose price is changed, while the quantity consumed of other goods remains the same. Since Sergei purchases all his products out of the same budget, a change in the price of one good can also have a range of effects, either positive or negative, on the quantity consumed of other goods.

In short, a higher price typically causes reduced consumption of the good in question, but it can affect the consumption of other goods as well.

### Try It
Visit this page in your course online to check your understanding.

### Watch It
We watched the first part of this video earlier when learning about budget constraints, and now we can watch the rest to see how the budget constraint graph changes with a change in budget or a change in the price of the goods. Visit this page in your course online to view this presentation.

### Try It
These questions allow you to get as much practice as you need, as you can click the link at the top of the first question (“Try another version of these questions”) to get a new set of questions. Practice until you feel comfortable doing the questions. Visit this page in your course online to practice before taking the quiz.

### Glossary

- **budget constraint**: all possible consumption combinations of goods that someone can afford, given the prices of goods, when all income is spent; the boundary of the opportunity set

- **income effect**: a higher price means that, in effect, the buying power of income has been reduced, even though actual income has not changed; always happens simultaneously with a substitution effect

- **substitution effect**: when a price changes, consumers have an incentive to consume less of the good with a relatively higher price and more of the good with a relatively low price.
THE FOUNDATIONS OF THE DEMAND CURVE

Learning Objectives

- Describe how demand curves are derived from consumer equilibrium

In this module, we’ve learned how consumers spend their limited income in order to maximize total utility. Each consumer choice problem yields a consumer equilibrium, showing the combination of goods and services an individual chooses to purchase with their budget, given the individual’s preferences and given the prices of the goods and services available.

Remember that a demand curve shows the relationship between price of a product and quantity demanded. While demand curves will appear somewhat different for each product – they may appear relatively steep or flat, straight or curved – demand curves slope down from left to right. So demand curves embody the law of demand: as the price increases, the quantity demanded decreases, and conversely, as the price decreases, the quantity demanded increases.

In this section, we explore the relationship between consumer equilibrium and demand curves.

The Foundations of Demand Curves

Figure 1 shows a budget constraint with a choice between housing and “everything else.” (Putting “everything else” on the vertical axis can be a useful approach in some cases, especially when the focus of the analysis is on one particular good.) We label the consumer equilibrium, \( M_0 \). This consumer equilibrium was found by applying the two rules used to maximize utility explained earlier.
Consider how changes in the price of a good affect the budget constraint and the consumer equilibrium. An increase in the price of housing will not change the quantity of “everything else” that the consumer is able to buy. If the consumer were to spend their entire budget on “everything else,” they could still afford the amount shown by the vertical axis of the budget constraint. But it will reduce the amount of housing the consumer can purchase. Thus, if the price of housing increases, the budget constraint will rotate towards the horizontal axis. This is shown in Figure 2 below.

The other three budget constraints represent successively higher prices for housing of $P_1$, $P_2$, and $P_3$. As the budget constraint rotates in, and in, and in again, we label the utility-maximizing choices $M_1$, $M_2$, and $M_3$, and the quantity demanded of housing falls from $Q_0$ to $Q_1$ to $Q_2$ to $Q_3$. 

---

**Figure 1.** A Choice between Housing and Everything Else. The points on the budget constraint show the combinations of Housing and a composite good of Everything Else that are affordable. The consumer equilibrium occurs at Point $M$. 

---
Figure 2. From Budget Constraint to Demand Curve. (a) As the price increases from $P_0$ to $P_1$ to $P_2$ to $P_3$, the budget constraint on the upper part of the diagram rotates clockwise. The utility-maximizing choice changes from $M_0$ to $M_1$ to $M_2$ to $M_3$. As a result, the quantity demanded of housing shifts from $Q_0$ to $Q_1$ to $Q_2$ to $Q_3$, ceteris paribus. (b) The demand curve graphs each combination of the price of housing and the quantity of housing demanded, ceteris paribus. The quantities of housing are the same at the points on both (a) and (b). Thus, the original price of housing ($P_0$) and the original quantity of housing ($Q_0$) appear on the demand curve as point $E_0$. The higher price of housing ($P_1$) and the corresponding lower quantity demanded of housing ($Q_1$) appear on the demand curve as point $E_1$.

So, as the price of housing rises, the budget constraint rotates clockwise (inward), and the quantity consumed of housing falls, ceteris paribus (meaning, with all other things being the same). This relationship—the price of housing rising from $P_0$ to $P_1$ to $P_2$ to $P_3$, while the quantity of housing demanded falls from $Q_0$ to $Q_1$ to $Q_2$ to $Q_3$—is graphed on the demand curve in Figure 2(b). Indeed, the vertical dashed lines stretching between the top and bottom of Figure 1 show that the quantity of housing demanded at each point is the same in both (a) and (b). Thus, the shape of a demand curve is ultimately determined by the underlying choices about maximizing utility subject to a budget constraint. And while economists may not be able to measure “utils,” they can certainly measure price and quantity demanded.

Applications in Government and Business

The budget constraint framework for making utility-maximizing choices offers a reminder that people can react to a change in price or income in a range of different ways. For example, in the winter months of 2005, costs for heating homes increased significantly in many parts of the country as prices for natural gas and electricity soared, due in large
part to the disruption caused by Hurricanes Katrina and Rita. Some people reacted by reducing the quantity demanded of energy; for example, by turning down the thermostats in their homes by a few degrees and wearing a heavier sweater inside. Even so, many home heating bills rose, so people adjusted their consumption in other ways, too. As you learned in the module on Elasticity, the short run demand for home heating is generally inelastic. Each household cut back on what it valued least on the margin; for some it might have been some dinners out, or a vacation, or postponing buying a new refrigerator or a new car. Indeed, sharply higher energy prices can have effects beyond the energy market, leading to a widespread reduction in purchasing throughout the rest of the economy.

A similar issue arises when the government imposes taxes on certain products, like it does on gasoline, cigarettes, and alcohol. Say that a tax on alcohol leads to a higher price at the liquor store, the higher price of alcohol causes the budget constraint to pivot left, and consumption of alcoholic beverages is likely to decrease. However, people may also react to the higher price of alcoholic beverages by cutting back on other purchases. For example, they might cut back on snacks at restaurants like chicken wings and nachos. It would be unwise to assume that the liquor industry is the only one affected by the tax on alcoholic beverages.

Does it make a difference who controls household income?

In the mid-1970s, the United Kingdom made an interesting policy change in its “child allowance” policy. This program provides a fixed amount of money per child to every family, regardless of family income. Traditionally, the child allowance had been distributed to families by withholding less in taxes from the paycheck of the family wage earner—typically the father in this time period. The new policy instead provided the child allowance as a cash payment to the mother. As a result of this change, households have the same level of income and face the same prices in the market, but the money is more likely to be in the purse of the mother than in the wallet of the father.

Should this change in policy alter household consumption patterns? Basic models of consumption decisions, of the sort examined in this module, assume that it does not matter whether the mother or the father receives the money, because both parents seek to maximize the utility of the family as a whole. In effect, this model assumes that everyone in the family has the same preferences.

In reality, the share of income controlled by the father or the mother does affect what the household consumes. When the mother controls a larger share of family income a number of studies, in the United Kingdom and in a wide variety of other countries, have found that the family tends to spend more on restaurant meals, child care, and women’s clothing, and less on alcohol and tobacco. As the mother controls a larger share of household resources, children’s health improves, too. These findings suggest that when providing assistance to poor families, in high-income countries and low-income countries alike, the monetary amount of assistance is not all that matters: it also matters which member of the family actually receives the money.

The budget constraint framework serves as a constant reminder to think about the full range of effects that can arise from changes in income or price, not just effects on the one product that might seem most immediately affected.

Try It

Visit this page in your course online to check your understanding.

INTRODUCTION TO INDIFFERENCE CURVES
What you'll learn to do: find consumer equilibrium using indifference curves and a budget constraint

Although numbers can be used to illustrate consumer preferences, economists don’t believe that we can objectively measure someone’s utility. In this section, you’ll learn an alternative way of identifying consumer equilibrium that only requires that you can rank preferences—that is, you can say whether you prefer Option A to Option B.

**INDIFFERENCE CURVE ANALYSIS**

**Learning Objectives**

- Describe the purpose, use, and shape of indifference curves
- Explain how one indifference curve differs from another
- Explain how to find the consumer equilibrium using indifference curves and a budget constraint
Economists use the vocabulary of maximizing utility to describe consumer choice. So far in the text, we have described the level of utility that a person receives in numerical terms. This section presents an alternative approach to describing personal preferences, called indifference curve analysis, which avoids the need for using numbers to measure utility. By setting aside the assumption of putting a numerical valuation on utility—an assumption that many students and economists find uncomfortably unrealistic—the indifference curve framework helps to clarify the logic of the underlying model.

What Is an Indifference Curve?

People cannot really put a numerical value on their level of satisfaction. However, they can, and do, identify what choices would give them more, or less, or the same amount of satisfaction. An indifference curve shows all combinations of goods that provide an equal level of utility or satisfaction.

For example, Figure 1 presents three indifference curves that represent Lilly’s preferences for the tradeoffs that she faces in her two main relaxation activities: eating doughnuts and reading paperback books. Each indifference curve (Ul, Um, and Uh) represents one level of utility. First we will explore the meaning of an individual indifference curve and then we will look at the relationship between different indifference curves.

**Figure 1. Lilly’s Indifference Curves.** Lilly would receive equal utility from all combinations of books and doughnuts on a given indifference curve. Any points on the highest indifference curve Uh, like F, provide greater utility than any points like A, B, C, and D on the middle indifference curve Um. Similarly, any points on the middle indifference curve Um provide greater utility than any points on the lowest indifference curve Ul.

**Watch It**

Watch the clip from this video carefully to see examples of indifference curves and what makes them useful. Each point on the indifference curves represents the same level of satisfaction.

Visit this page in your course online to view this presentation.

The Shape of an Indifference Curve
The indifference curve $U_m$ has four points labeled on it: $A$, $B$, $C$, and $D$ (see Figure 1). Since an indifference curve represents a set of choices that have the same level of utility, Lilly must receive an equal amount of utility, judged according to her personal preferences, from two books and 120 doughnuts (point A), from three books and 84 doughnuts (point B) from 11 books and 40 doughnuts (point C) or from 12 books and 35 doughnuts (point D). She would also receive the same utility from any of the unlabeled intermediate points along this indifference curve.

Indifference curves have a roughly similar shape in two ways: 1) they are downward sloping from left to right; 2) they are convex with respect to the origin. In other words, they are steeper on the left and flatter on the right. The downward slope of the indifference curve means that Lilly must trade off less of one good to get more of the other, while holding utility constant. For example, points $A$ and $B$ sit on the same indifference curve $U_m$, which means that they provide Lilly with the same level of utility. Thus, the marginal utility that Lilly would gain from, say, increasing her consumption of books from two to three must be equal to the marginal utility that she would lose if her consumption of doughnuts was cut from 120 to 84—so that her overall utility remains unchanged between points $A$ and $B$. Indeed, the slope along an indifference curve is the marginal rate of substitution, which is the rate at which a person is willing to trade one good for another so that utility will remain the same.

Indifference curves like $U_m$ are steeper on the left and flatter on the right. The reason behind this shape involves diminishing marginal utility—the notion that as a person consumes more of a good, the marginal utility from each additional unit becomes lower. Compare two different choices between points that all provide Lilly an equal amount of utility along the indifference curve $U_m$: the choice between $A$ and $B$, and between $C$ and $D$. In both choices, Lilly consumes one more book, but between $A$ and $B$ her consumption of doughnuts falls by 36 (from 120 to 84) and between $C$ and $D$ it falls by only five (from 40 to 35). The reason for this difference is that points $A$ and $C$ are different starting points, and thus have different implications for marginal utility. At point $A$, Lilly has few books and many doughnuts. Thus, her marginal utility from an extra book will be relatively high while the marginal utility of additional doughnuts is relatively low—so on the margin, it will take a relatively large number of doughnuts to offset the utility from the marginal book. At point $C$, however, Lilly has many books and few doughnuts. From this starting point, her marginal utility gained from extra books will be relatively low, while the marginal utility lost from additional doughnuts would be relatively high—so on the margin, it will take a relatively smaller number of doughnuts to offset the change of one marginal book. In short, the slope of the indifference curve changes because the marginal rate of substitution—that is, the quantity of one good that would be traded for the other good to keep utility constant—also changes, as a result of diminishing marginal utility of both goods.

**Try It**

Visit this page in your course online to check your understanding.

**The Field of Indifference Curves**

Each indifference curve represents the choices that provide a single level of utility. Every level of utility will have its own indifference curve. Thus, Lilly’s preferences will include an infinite number of indifference curves lying nested together on the diagram—even though only three of the indifference curves, representing three levels of utility, appear in Figure 1. In other words, an infinite number of indifference curves are not drawn on this diagram—but you should remember that they exist.

Higher indifference curves represent a greater level of utility than lower ones. In Figure 1, indifference curve $U_l$ can be thought of as a “low” level of utility, while $U_m$ is a “medium” level of utility and $U_h$ is a “high” level of utility. All of the choices on indifference curve $U_h$ are preferred to all of the choices on indifference curve $U_m$, which in turn are preferred to all of the choices on $U_l$.

To understand why higher indifference curves are preferred to lower ones, compare point $B$ on indifference curve $U_m$ to point $F$ on indifference curve $U_h$. Point $F$ has greater consumption of both books (five to three) and doughnuts (100 to 84), so point $F$ is clearly preferable to point $B$. Given the definition of an indifference curve—that all the points on the curve have the same level of utility—if point $F$ on indifference curve $U_h$ is preferred to point $B$ on indifference curve $U_m$, then it must be true that all points on indifference curve $U_h$ have a higher level of utility than all points on $U_m$. More generally, for any point on a lower indifference curve, like $U_l$, you can identify a point on a higher indifference curve like $U_m$ or $U_h$ that has a higher consumption of both goods. Since one point on the higher indifference curve is preferred to one point on the lower curve, and since all the points on a given indifference curve have the same level of utility, it must be true that all points on higher indifference curves have greater utility than all points on lower indifference curves.
These arguments about the shapes of indifference curves and about higher or lower levels of utility do not require any numerical estimates of utility, either by the individual or by anyone else. They are only based on the assumptions that when people have less of one good they need more of another good to make up for it, if they are keeping the same level of utility, and that as people have more of a good, the marginal utility they receive from additional units of that good will diminish. Given these gentle assumptions, a field of indifference curves can be mapped out to describe the preferences of any individual.

Try It
Visit this page in your course online to check your understanding.

The Individuality of Indifference Curves

Each person determines his or her own preferences and utility. Thus, while indifference curves have the same general shape—they slope down, and the slope is steeper on the left and flatter on the right—the specific shape of indifference curves can be different for every person. Figure 1, for example, applies only to Lilly’s preferences. Indifference curves for other people would probably travel through different points.

Utility-Maximizing with Indifference Curves

People seek the highest level of utility, which means that they wish to be on the highest possible indifference curve. However, people are limited by their budget constraints, which show what tradeoffs are actually possible.

Maximizing Utility at the Highest Indifference Curve

Return to the situation of Lilly’s choice between paperback books and doughnuts. Say that books cost $6, doughnuts are 50 cents each, and that Lilly has $60 to spend. This information provides the basis for the budget line shown in Figure 1. Along with the budget line are shown the three indifference curves from Figure 1. What is Lilly’s utility-maximizing choice? Several possibilities are identified in the diagram.
The choice of F with five books and 100 doughnuts is highly desirable, since it is on the highest indifference curve Uh of those shown in the diagram. However, it is not affordable given Lilly’s budget constraint. The choice of H with three books and 70 doughnuts on indifference curve Ul is a wasteful choice, since it is inside Lilly’s budget set, and as a utility-maximizer, Lilly will always prefer a choice on the budget constraint itself. Choices B and G are both on the opportunity set. However, choice G of six books and 48 doughnuts is on lower indifference curve Ul than choice B of three books and 84 doughnuts, which is on the indifference curve Um. If Lilly were to start at choice G, and then thought about whether the marginal utility she was deriving from doughnuts and books, she would decide that some additional doughnuts and fewer books would make her happier—which would cause her to move toward her preferred choice B. Given the combination of Lilly’s personal preferences, as identified by her indifference curves, and Lilly’s opportunity set, which is determined by prices and income, B will be her utility-maximizing choice.

The highest achievable indifference curve touches the budget constraint at a single point of tangency. Since an infinite number of indifference curves exist, even if only a few of them are drawn on any given diagram, there will always exist one indifference curve that touches the budget line at a single point of tangency. All higher indifference curves, like Uh, will be completely above the budget line and, although the choices on that indifference curve would provide higher utility, they are not affordable given the budget set. All lower indifference curves, like Ul, will cross the budget line in two separate places. When one indifference curve crosses the budget line in two places, however, there will be another, higher, attainable indifference curve sitting above it that touches the budget line at only one point of tangency.
Watch this video to apply what you’ve learned about using indifference curves to find consumer equilibrium. Watch this video online: https://youtu.be/MXIgp-P-FeY

Try It

These questions allow you to get as much practice as you need, as you can click the link at the top of the first question (“Try another version of these questions”) to get a new set of questions. Practice until you feel comfortable doing the questions.
Visit this page in your course online to practice before taking the quiz.

INTRODUCTION TO BEHAVIORAL ECONOMICS

What you’ll learn to do: describe the behavioral economics approach to understanding decision making

If a friendly stranger handed you $20, how would you feel? What if you lost $20 the next week? Do you think that you’d feel sad about it, or would just be happy that you broke even? In this section, you’ll see that most people would still be sad about the $20 loss. Economic decisions are not always logical. Read on to find out why!
BEHAVIORAL ECONOMICS: AN ALTERNATIVE VIEWPOINT

Learning Objectives

- Describe behavioral economics
- Give examples of irrational decision-making

As we know, people sometimes make decisions that seem “irrational” and not in their own best interest. People’s decisions can seem inconsistent from one day to the next and they even deliberately ignore ways to save money or time. The traditional economic models assume rationality, which means that people take all available information and make consistent and informed decisions that are in their best interest. (In fact, economics professors often delight in pointing out so-called “irrational behavior” each semester to their new students, and present economics as a way to become more rational.)

However, a new group of economists, known as behavioral economists, argue that the traditional method omits something important: people’s state of mind. For example, one can think differently about money if one is feeling revenge, optimism, or loss. These are not necessarily irrational states of mind, but part of a range of emotions that can affect anyone on a given day. In addition, actions under these conditions are predictable, if one better understands the underlying environment. Behavioral economics seeks to enrich our understanding of decision-making by integrating the insights of psychology into economics. It does this by investigating how given dollar amounts can mean different things to individuals depending on the situation. This can lead to decisions that appear outwardly inconsistent, or irrational, to the outside observer.

The way the mind works, according to this view, may seem inconsistent to traditional economists but is actually far more complex than an unemotional cost-benefit adding machine. For example, a traditional economist would say that if you lost a $10 bill today, and also received an extra $10 in your paycheck, you should feel perfectly neutral. After all, –$10 + $10 = $0. You are the same financially as you were before. However, behavioral economists have conducted research that shows many people will feel some negative emotion, such as anger or frustration, after those two things happen. We tend to focus more on the loss than the gain. We call this loss aversion, where a $1 loss pains us 2.25 times more than a $1 gain helps us, according to the economists Daniel Kahneman and Amos Tversky in a famous 1979 article in the journal *Econometrica*. This insight has implications for investing, as people tend to “overplay” the stock market by reacting more to losses than to gains. This behavior looks irrational to traditional economists, but is consistent once we understand better how the mind works, these economists argue.

Watch It

Economic models are useful tools in predicting behavior, but people sometimes act irrationally and differently than the ways these models suggest. One reason for this is bounded irrationality, which means that consumers face limits on their information, time, or abilities that prevent them from seeking out the best possible outcomes. Watch this video to
behavioral economics: see examples of how economists and psychologists have done research on things like the ultimatum game, nudges, and loss aversion, to help explain consumer irrationality.  
Watch this video online: https://youtu.be/dqxQ3E1bubI

Irrational Consumer Behavior

Traditional economists also assume human beings have complete self control, but, for instance, people will buy cigarettes by the pack instead of the carton even though the carton saves them money, to keep usage down. They purchase locks for their refrigerators and overpay on taxes to force themselves to save. In other words, we protect ourselves from our worst temptations but pay a price to do so. One way behavioral economists are responding to this is by establishing ways for people to keep themselves free of these temptations. This includes what we call “nudges” toward more rational behavior rather than mandatory regulations from government. For example, up to 20 percent of new employees do not enroll in retirement savings plans immediately, because of procrastination or feeling overwhelmed by the different choices. Some companies are now moving to a new system, where employees are automatically enrolled unless they “opt out.” Almost no-one opts out in this program and employees begin saving at the early years, which are most critical for retirement.

Another area that seems illogical is the idea of mental accounting, or putting dollars in different mental categories where they take different values. Economists typically consider dollars to be fungible, or having equal value to the individual, regardless of the situation.

You might, for instance, think of the $25 you found in the street differently from the $25 you earned from three hours working in a fast food restaurant. The street money might well be treated as “mad money” with little rational regard to getting the best value. This is in one sense strange, since it is still equivalent to three hours of hard work in the restaurant. Yet the “easy come-easy go” mentality replaces the rational economizer because of the situation, or context, in which the money was attained.

In another example of mental accounting that seems inconsistent to a traditional economist, a person could carry a credit card debt of $1,000 that has a 15% yearly interest cost, and simultaneously have a $2,000 savings account that pays only 2% per year. That means she pays $150 a year to the credit card company, while collecting only $40 annually in bank interest, so she loses $130 a year. That doesn’t seem wise.

The “rational” decision would be to pay off the debt, since a $1,000 savings account with $0 in debt is the equivalent net worth, and she would now net $20 per year. But curiously, it is not uncommon for people to ignore this advice, since they will treat a loss to their savings account as higher than the benefit of paying off their credit card. The dollars are not being treated as fungible so it looks irrational to traditional economists.

Which view is right, the behavioral economists’ or the traditional view? Both have their advantages, but behavioral economists have at least shed a light on trying to describe and explain behavior that has historically been dismissed as irrational. If most of us are engaged in some “irrational behavior,” perhaps there are deeper underlying reasons for this behavior in the first place.

Watch It

Although economists would love to assume that all people think rationally about their financial choices, events like the financial crisis led behavioral economists to look at how people actually make decisions in financial trading, instead of how they say they do, or even how they should. This video explains how human psychology played a large role in the financial crisis of 2008.

Watch this video online: https://youtu.be/GZB616a7eb0

glossary

behavioral economics: a branch of economics that seeks to enrich the understanding of decision-making by integrating the insights of psychology and by investigating how given dollar amounts can mean different things to individuals depending on the situation
fungible: the idea that units of a good, such as dollars, ounces of gold, or barrels of oil are capable of mutual substitution with each other and carry equal value to the individual

PUTTING IT TOGETHER: UTILITY

The key underlying principle in this module was getting the biggest bang for the buck. This principle will be used over and over again in different contexts in this text, and the results will sometimes surprise you.

Making Choices During the Recession

Remember the question posed at the beginning of this module—in what category did consumers worldwide increase their spending during the recession? The answer is higher education. According to the United Nations Educational, Scientific, and Cultural Organization (UNESCO), enrollment in colleges and universities rose one-third in China and almost two-thirds in Saudi Arabia, nearly doubled in Pakistan, tripled in Uganda, and surged by three million—18 percent—in the United States. Why were consumers willing to spend on education during lean times? Both individuals and countries view higher education as the way to prosperity. Many feel that increased earnings are a significant benefit of attending college.

Bureau of Labor Statistics data from May 2012 supports this view, as shown in Figure 3. They show a positive correlation between earnings and education. The data also indicate that unemployment rates fall with higher levels of education and training.
Figure 1. The Impact of Education on Earnings and Unemployment Rates, 2012. Those with the highest degrees in 2012 had substantially lower unemployment rates whereas those with the least formal education suffered from the highest unemployment rates. The national median average weekly income was $815, and the nation unemployment average in 2012 was 6.8%. (Source: Bureau of Labor Statistics, May 22, 2013)

Remember the other example we posed at the beginning of the module—what would be the better strategy for maximizing your GPA? To keep it simple, let us suppose you are only taking two courses, an “easy” course, and a “hard” course, where the difficulty is defined by how much time and effort it takes to earn a given grade. Is Principles of Microeconomics an easy course or a hard one for you? What about Biology? You get to decide which of your courses falls into each category. Using this example, many people think that it makes the most sense to spend the most time and effort on the hard course. After all, a hard course requires more time to learn, right?

That’s true, but if you think like an economist, you’ll see that to maximize your GPA, given a limited amount of study time, it makes more sense to start with the course where your study time will have the most impact on your grades, the biggest bang for the buck. In other words, you should start with the easy course and quite possibly spend more time on it, to assure yourself of an A.

This is no different than choosing to spend more of your budget on the product that gives you the most marginal utility per dollar spent. It both cases, you make the most of your scarce resources—budget dollars in the consumption case and study hours in the GPA case.

As you proceed through the rest of this text, look for opportunities to apply the biggest bang for the buck principle.
MODULE 7: PRODUCTION AND COSTS

WHY IT MATTERS: PRODUCTION AND COSTS

Why analyze the relationship between inputs used in production, and the resulting outputs and costs?

Figure 1. “Bread Collage”. Amazon’s Front Door by Robert Scoble, CC-BY. Tesla Autobots by Steve Jurvetson, CC-BY. Road Building by Elvert Barnes, CC-BY-SA. Baking Bread by Christian Guthier, CC-BY. House framing by Ryan, CC-BY-SA.

We are changing gears in this module, which is the first of five dealing with the theory of the firm. The theory of the firm seeks to understand what motivates firms to make the operating decisions that they do. This module, focusing on production and costs (the cost of producing a given quantity of output), provides the necessary background information to do that.

This module has a lot of detail, tables and graphs, but there are really two main key takeaways:

1. There are several different ways to look at production and costs, just like there are several different ways to look at a student’s learning in a course. Students can be asked to write an essay or do a class presentation on a topic. These are two ways of trying to assess the same thing, student learning.
2. There is an inverse relationship between production and costs. The harder it is to produce something, for example, the more labor it takes, the higher the cost of producing it, and vice versa.
Watch It

As production of cars increased, the cost of producing them (and thus the price charged to customers) decreased dramatically. Watch this video to learn more.
Watch this video online: https://youtu.be/iD6mOxILbjQ

This module is very detail-oriented. The graphs you will encounter are important because they'll help you keep the different concepts, like production versus costs or different types of cost, straight. Keep track of what’s on the horizontal and vertical axis and you'll know what's going on. If employment (or L) is on the horizontal axis, you’re looking at some kind of a production curve. If output (or Q) is on the horizontal axis, you’re looking at some kind of cost curve.

Students sometimes misunderstand the concept of technology. In an economics context, technology does not mean high tech, or having to do with computers or the internet. Rather, technology means the way, the process by which inputs are used to produce outputs. A different process means a different technology.

Here are some questions to think about: How does an improvement in technology affect production and costs? How, for example, would it affect the graphs? We will answer this at the end of the module, once you have the tools.

One more thing: do not confuse cost with price. This module is about firms. From the firm's perspective, cost is what they pay for the inputs necessary to produce the product. Price is what the firm receives for selling the product. Thus, cost is a negative and price is positive; they are not the same thing.

INTRODUCTION TO PRODUCTION

What you’ll learn to do: explain production and the production function

Figure 1. Different types of production: a steel mill and a wheat field.
Businesses engage in production. What does that mean exactly? What is involved? How is steel production different from growing wheat? In this section, we will begin to learn about the behavior of firms, how they make production decisions, and how production costs depend on a firm’s production function.

WHAT IS PRODUCTION?

Learning Objectives

- Describe and differentiate between the types of inputs in the production process

As we said earlier, this module is the first of several modules that explore the theory of the firm. Let’s define what is meant by the firm. A firm (or business) combines inputs of labor, capital, land, and raw or finished component materials to produce outputs. If the firm is successful, the outputs are more valuable than the inputs. This activity of production goes beyond manufacturing (i.e., making things). It includes any process or service that creates value, including transportation, distribution, wholesale and retail sales. Production involves a number of important decisions that define the behavior of firms. These decisions include, but are not limited to:

- What product or products should the firm produce?
- How should the products be produced (i.e., what production process should be used)?
- How much output should the firm produce?
- What price should the firm charge for its products?
- How much labor should the firm employ?

The answers to these questions depend on the production and cost conditions facing each firm, which is the subject of this module. The answers also depend on the structure of the market for the product(s) in question. Market structure is a multidimensional concept that involves how competitive an industry is. It is defined by questions such as these:

- How much market power does each firm in the industry possess?
- How similar is each firm’s product to the products of other firms in the industry?
- How difficult is it for new firms to enter the industry?
- Do firms compete on the basis of price, advertising, or other product differences?

Figure 1 illustrates the range of different market structures, which we will explore in detail in later modules.
Factors of Production

In this module, we want to explore the relationship between the quantity of output a firm produces, and the cost of producing that output. The cost of the product depends on how many inputs (or factors of production) are required to produce the product and what those inputs cost. We can determine the costs by looking at the firm’s production function, which we will explore in detail in the next section.

Production is the process (or processes) a firm uses to transform inputs (e.g. labor, capital, raw materials) into outputs, i.e. the goods or services the firm wishes to sell. Consider pizza making. The pizzaiolo (pizza maker) takes flour, water, and yeast to make dough. Similarly, the pizzaiolo may take tomatoes, spices, and water to make pizza sauce. He or she rolls out the dough, brushes on the pizza sauce, and adds cheese and other toppings. The pizzaiolo uses a peel—the shovel-like wooden tool—to put the pizza into the oven to cook. Once baked, the pizza goes into a box (if it's for takeout) and the customer pays for the good. What are the inputs (or factors of production) in the production process for this pizza?

Economists divide factors of production into several categories:

- **Natural Resources** (Land and Raw Materials) – The ingredients for the pizza are raw materials. These include the flour, yeast, and water for the dough, the tomatoes, herbs, and water for the sauce, the cheese, and the toppings. If the pizza place uses a wood-burning oven, we would include the wood as a raw material. If the establishment heats the oven with natural gas, we would count this as a raw material. Don’t forget electricity for lights. If, instead of pizza, we were looking at an agricultural product, like wheat, we would include the land the farmer used for crops here.
**Factors of Production (or Inputs):**

- **Labor:** When we talk about production, labor means human effort, both physical and mental. The pizzaiolo was the primary example of labor here. He or she needs to be strong enough to roll out the dough and to insert and retrieve the pizza from the oven, but he or she also needs to know how to make the pizza, how long it cooks in the oven and a myriad of other aspects of pizza-making. The business may also have one or more people to work the counter, take orders, and receive payment.

- **Capital:** When economists use the term capital, they do not mean financial capital (money); rather, they mean physical capital, the machines, equipment, and buildings that one uses to produce the product. In the case of pizza, the capital includes the peel, the oven, the building, and any other necessary equipment (for example, tables and chairs).

- **Technology:** Technology refers to the process or processes for producing the product. How does the pizzaiolo combine ingredients to make pizza? How hot should the oven be? How long should the pizza cook? What is the best oven to use? Gas or wood burning? Should the restaurant make its own dough, sauce, cheese, toppings, or should it buy them?

- **Entrepreneurship:** Production involves many decisions and much knowledge, even for something as simple as pizza. Who makes those decisions? Ultimately, it is the entrepreneur, the person who creates the business, whose idea it is to combine the inputs to produce the outputs.

The cost of producing pizza (or any output) depends on the amount of labor capital, raw materials, and other inputs required and the price of each input to the entrepreneur.

---

**Try It**

Visit this page in your course online to view this presentation.

---

**Glossary**

- **Factors of production (or inputs):** resources that firms use to produce their products, for example, labor and capital
- **Firm:** an organization that combines inputs of labor, capital, land, and raw or finished component materials to produce outputs.
- **Production:** the process of combining inputs to produce outputs, ideally of a value greater than the value of the inputs

---

**The Production Function**

**Learning Objectives**

- Explain the concept of a production function
- Differentiate between fixed and variable inputs
• Differentiate between total and marginal product
• Describe diminishing marginal productivity

We can summarize the ideas so far in terms of a production function, a mathematical expression or equation that explains the relationship between a firm’s inputs and its outputs:

\[ Q = f [NR,L,K,t,E] \]

A production is purely an engineering concept. If you plug in the amount of labor, capital and other inputs the firm is using, the production function tells how much output will be produced by those inputs. Production functions are specific to the product. Different products have different production functions. The amount of labor a farmer uses to produce a bushel of corn is likely different than that required to produce an automobile. Firms in the same industry may have somewhat different production functions, since each firm may produce a little differently. One pizza restaurant may make its own dough and sauce, while another may buy those pre-made. A sit-down pizza restaurant probably uses more labor (to handle table service) than a purely take-out restaurant. We can describe inputs as either fixed or variable.

**Fixed inputs** are those that can’t easily be increased or decreased in a short period of time. In the pizza example, the building is a fixed input. Once the entrepreneur signs the lease, he or she is stuck in the building until the lease expires. Fixed inputs define the firm’s maximum output capacity. This is analogous to the potential real GDP shown by society’s production possibilities curve, i.e. the maximum quantities of outputs a society can produce at a given time with its available resources. Fixed inputs do not change as output changes.

**Variable inputs** are those that can easily be increased or decreased in a short period of time. The pizzaiolo can order more ingredients with a phone call, so ingredients would be variable inputs. The owner could hire a new person to work the counter pretty quickly as well. Variable inputs increase or decrease as output changes.

Economists often use a short-hand form for the production function:

\[ Q = f [L,K] \]

where L represents all the variable inputs, and K represents all the fixed inputs.

**Try It**

Visit this page in your course online to check your understanding.

Economists also differentiate between short and long run production. The **short run** is the period of time during which at least some factors of production are fixed. During the period of the pizza restaurant lease, the pizza restaurant is operating in the short run, because it is limited to using the current building—the owner can’t choose a larger or smaller building. The **long run** is the period of time during which all factors are variable. Once the lease expires for the pizza restaurant, the shop owner can move to a larger or smaller place.

Note that there is another important distinction between fixed and variable inputs. In the short run, since the firm’s fixed inputs are fixed, the only way to vary a firm’s output is by changing its variable inputs. Let’s explore production in the short run using a specific example: tree cutting (for lumber) with a two-person crosscut saw.
Since by definition capital is fixed in the short run, our production function becomes

\[ Q = f \left[L, \bar{K}\right] \text{ or } Q = f [L] \]

This equation simply indicates that since capital is fixed, then changing the amount of output (e.g. trees cut down per day) depends only on changing the amount of labor employed (e.g. number of lumberjacks working). We can express this production function numerically as Table 1 below shows. You can also see it graphically in Figure 2a.

<table>
<thead>
<tr>
<th># Lumberjacks</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td># Trees (TP)</td>
<td>4</td>
<td>10</td>
<td>12</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>MP</td>
<td>4</td>
<td>6</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
Figure 2. Total Product and Marginal Product Curves. The short run total product for trees (top) shows the amount of output produced with fixed capital. In this example, one lumberjack using a two-person saw can cut down four trees in an hour. Three lumberjacks using a two-person saw can cut down twelve trees in an hour. The marginal product for trees (bottom) shows the additional output created by one more lumberjack.

Note that we have introduced some new language. We also call Output (Q) Total Product (TP), which means the amount of output produced with a given amount of labor and a fixed amount of capital. In this example, one lumberjack using a two-person saw can cut down four trees in an hour. Two lumberjacks using a two-person saw can cut down ten trees in an hour.

Try It

Visit this page in your course online to check your understanding.
We should also introduce a critical concept: marginal product. **Marginal product** is the additional output of one more worker. Mathematically, Marginal Product is the change in total product divided by the change in labor: 

\[
MP = \frac{\Delta TP}{\Delta L}
\]

In the table above, since 0 workers produce 0 trees, the marginal product of the first worker is four trees per day, but the marginal product of the second worker is six trees per day. Why might that be the case? It's because of the nature of the capital the workers are using. A two-person saw works much better with two persons than with one. Suppose we add a third lumberjack to the story. What will that person's marginal product be? What will that person contribute to the team? Perhaps he or she can oil the saw's teeth to keep it sawing smoothly or he or she could bring water to the two people sawing.

What you see in the table is a critically important conclusion about production in the short run: it may be that as we add workers, the marginal product increases at first, but sooner or later additional workers will have decreasing marginal product. In fact, there may eventually be no effect or a negative effect on output. This is called the **Law of Diminishing Marginal Product** and it's a characteristic of production in the short run. Diminishing marginal productivity is very similar to the concept of diminishing marginal utility that we learned about in the chapter on consumer choice. Both concepts are examples of the more general concept of diminishing marginal returns. Why does diminishing marginal productivity occur? It's because of fixed capital. We will see this more clearly when we discuss production in the long run.

We can show these concepts graphically, as you can see in Figure 2 above. Figure 3 shows the more general cases of total product and marginal product curves.

![Figure 3. Total Product and Marginal Product Curves.](image)

The top graph shows the general shape of a total product curve, with total product initially increasing, then tapering off due to the law of diminishing marginal product. The bottom graph shows how marginal product falls with additional labor.
INTRODUCTION TO COSTS IN THE SHORT RUN
What you’ll learn to do: calculate, graph and understand production costs in the short run

It’s obvious that a firm’s total revenue must exceed total costs if it wants to make a profit. In this section, you’ll see why it’s helpful for firms to break down and examine their costs in different ways. We will look at fixed versus variable costs, average versus marginal costs, and more. The purpose of these different concepts is to give the firm a better picture of how they can become more profitable.

COSTS AND PROFIT

Learning Objectives

- Explain the difference between explicit costs and implicit costs
- Calculate accounting and economic profit

Explicit and Implicit Costs, and Accounting and Economic Profit

Private enterprise, the ownership of businesses by private individuals, is a hallmark of the U.S. economy. When people think of businesses, often giants like Wal-Mart, Microsoft, or General Motors come to mind. But firms come in all sizes, as shown in Table 1. The vast majority of American firms have fewer than 20 employees. As of 2010, the U.S. Census
Bureau counted 5.7 million firms with employees in the U.S. economy. Slightly less than half of all the workers in private firms are at the 17,000 large firms, meaning those firms each employ at least 500 workers. Another 35% of workers in the U.S. economy are at firms with fewer than 100 workers. These small-scale businesses include everything from dentists and lawyers to businesses that mow lawns or clean houses. Indeed, Table 1 does not include a separate category for the millions of small “non-employer” businesses where a single owner or a few partners are not officially paid wages or a salary, but simply receive whatever they can earn.

Table 1. Range in Size of U.S. Firms (Source: U.S. Census, 2010 www.census.gov)

<table>
<thead>
<tr>
<th>Number of Employees</th>
<th>Firms (% of total firms)</th>
<th>Number of Paid Employees (% of total employment)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>5,734,538</td>
<td>112.0 million</td>
</tr>
<tr>
<td>0–9</td>
<td>4,543,315 (79.2%)</td>
<td>12.3 million (11.0%)</td>
</tr>
<tr>
<td>10–19</td>
<td>617,089 (10.8%)</td>
<td>8.3 million (7.4%)</td>
</tr>
<tr>
<td>20–99</td>
<td>475,125 (8.3%)</td>
<td>18.6 million (16.6%)</td>
</tr>
<tr>
<td>100–499</td>
<td>81,773 (1.4%)</td>
<td>15.9 million (14.2%)</td>
</tr>
<tr>
<td>500 or more</td>
<td>17,236 (0.30%)</td>
<td>50.9 million (49.8%)</td>
</tr>
</tbody>
</table>

Each of these businesses, regardless of size or complexity, tries to earn a profit:

Profit = Total Revenue – Total Cost

Total revenue is the income brought into the firm from selling its products. It is calculated by multiplying the price of the product times the quantity of output sold:

Total Revenue = Price x Quantity

As we study the theory of the firm, it will become clear that a firm’s revenue depends on the demand for the firm’s products.

We can distinguish between two types of cost: explicit and implicit. **Explicit costs** are out-of-pocket costs, that is, payments that are actually made. Wages that a firm pays its employees or rent that a firm pays for its office are explicit costs. **Implicit costs** are more subtle, but just as important. They represent the opportunity cost of using resources already owned by the firm. Often for small businesses, they are resources contributed by the owners; for example, working in the business while not getting a formal salary, or using the ground floor of a home as a retail store. Implicit costs also include the depreciation of goods, materials, and equipment that are necessary for a company to operate.

### Try It

Visit this page in your course online to check your understanding.

These two definitions of cost are important for distinguishing between two conceptions of profit, accounting profit and economic profit. **Accounting profit** is a cash concept. It means total revenue minus explicit costs—the difference between dollars brought in and dollars paid out. **Economic profit** is total revenue minus total cost, including both explicit and implicit costs. The difference is important because even though a business pays income taxes based on its accounting profit, whether or not it is economically successful depends on its economic profit.

### Watch It

Watch this video online: [https://youtu.be/a0nUWrnuUdo](https://youtu.be/a0nUWrnuUdo)

### CALCULATING IMPLICIT COSTS
Consider the following example. Fred currently works for a corporate law firm. He is considering opening his own legal practice, where he expects to earn $200,000 per year once he gets established. To run his own firm, he would need an office and a law clerk. He has found the perfect office, which rents for $50,000 per year. A law clerk could be hired for $35,000 per year. If these figures are accurate, would Fred's legal practice be profitable?

**Step 1.** First you have to calculate the costs. You can take what you know about explicit costs and total them:

<table>
<thead>
<tr>
<th>Office rental</th>
<th>$50,000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Law clerk’s salary</td>
<td>+ $35,000</td>
</tr>
<tr>
<td>Total explicit costs</td>
<td>$85,000</td>
</tr>
</tbody>
</table>

**Step 2.** Subtracting the explicit costs from the revenue gives you the accounting profit.

<table>
<thead>
<tr>
<th>Revenues</th>
<th>$200,000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Explicit costs</td>
<td>– $85,000</td>
</tr>
<tr>
<td>Accounting profit</td>
<td>$115,000</td>
</tr>
</tbody>
</table>

But these calculations consider only the explicit costs. To open his own practice, Fred would have to quit his current job, where he is earning an annual salary of $125,000. This would be an implicit cost of opening his own firm.

**Step 3.** You need to subtract both the explicit and implicit costs to determine the true economic profit. The equation is:

\[
\text{Economic Profit} = \text{Total Revenues} - \text{Explicit Costs} - \text{Implicit Costs}
\]

Now let's plug in Fred's figures to the true economic profit equation:

\[
\text{Economic Profit} = 200,000 - 85,000 - 125,000 = -10,000 \text{ per year}
\]

Fred would be losing $10,000 per year. That does not mean he would not want to open his own business, but it does mean he would be earning $10,000 less than if he worked for the corporate firm.

Implicit costs can include other things as well. Maybe Fred values his leisure time, and starting his own firm would require him to put in more hours than at the corporate firm. In this case, the lost leisure would also be an implicit cost that would subtract from economic profits.

Try It

Visit this page in your course online to view this presentation.

Now that we have an idea about the different types of costs, let's look at cost structures. A firm's cost structure in the long run may be different from that in the short run. We turn to that distinction in the next section.

glossary

**accounting profit** total revenues minus explicit costs, including depreciation

**economic profit** total revenues minus total costs (explicit plus implicit costs)

**explicit costs**: out-of-pocket costs for a firm, for example, payments for wages and salaries, rent, or materials

**firm**: an organization that combines inputs of labor, capital, land, and raw or finished component materials to produce outputs.

**implicit costs**: opportunity cost of resources already owned by the firm and used in business, for example, expanding a factory onto land already owned

**private enterprise**: the ownership of businesses by private individuals

**production**: 
costs in the short run

learning objectives

- describe the relationship between production and costs, including average and marginal costs
- analyze short-run costs in terms of fixed cost and variable cost

we've explained that a firm’s total cost of production depends on the quantities of inputs the firm uses to produce its output and the cost of those inputs to the firm. the firm’s production function tells us how much output the firm will produce with given amounts of inputs. a production function can be expressed mathematically as

\[ Q = f\left[L, K\right] \]

where \( Q \) is the firm’s output, \( L \) is the amount of labor employed, and \( K \) is the amount of fixed capital.

suppose we think about the production function backwards:

\[ L = g\left[Q, K\right], \]

where the \( g \) just means the function \( f \) in reverse. this equation tells us how much labor we need to produce a given level of output, with the fixed capital stock we have. if we knew the cost of labor and capital, we could then compute the total cost of producing any level of output. it is to this that we next turn.

for every factor of production (or input), there is an associated factor payment. **factor payments** are what the firm pays for the use of the factors of production. from the firm’s perspective, factor payments are costs. from the owner of each factor’s perspective, factor payments are income. factor payments include:

- raw materials prices for raw materials
- rent for land or buildings
- wages and salaries for labor
- interest and dividends for the use of financial capital (loans and equity investments)
- profit for entrepreneurship. profit is the residual, what’s left over from revenues after the firm pays all the other costs. while it may seem odd to treat profit as a “cost”, it is the payment that goes from total revenues to entrepreneurs or taking the risk of starting a business. you can see this correspondence between factors of production and factor payments in the inside loop of the circular flow diagram in figure 1.

we now have all the information necessary to determine a firm’s costs.
A cost function is a mathematical equation that shows the cost of producing different levels of output. Table 1 gives an example, which shows the cost of producing different quantities of widgets.

<table>
<thead>
<tr>
<th>Q</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cost</td>
<td>$32.50</td>
<td>$44</td>
<td>$52</td>
<td>$90</td>
</tr>
</tbody>
</table>

What we observe is that the cost increases as the firm produces higher quantities of output. This is pretty intuitive, since producing more output requires greater quantities of inputs, which cost more dollars to acquire.

What is the origin of these cost figures? They come from the production function and the factor payments. Suppose the production function for widgets is as shown in Table 2:

<table>
<thead>
<tr>
<th>Workers (L)</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>3.25</th>
<th>4.4</th>
<th>5.2</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Widgets (Q)</td>
<td>0.2</td>
<td>0.4</td>
<td>0.8</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>3.5</td>
<td>3.8</td>
<td>3.95</td>
<td>4</td>
</tr>
</tbody>
</table>

We can use the information from the production function to determine production costs. What we need to know is how many workers are required to produce any quantity of output. If we flip the order of the rows, we “invert” the production function so it shows \( L = g(Q) \).

Now focus on the whole number quantities of output. We'll eliminate the fractions (or partial widgets) from the table:

<table>
<thead>
<tr>
<th>Widgets (Q)</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>3.25</th>
<th>4.4</th>
<th>5.2</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Workers (L)</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>3.25</td>
<td>4.4</td>
<td>5.2</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
</tr>
</tbody>
</table>
Suppose widget workers receive $10 per hour. Multiplying the Workers row by $10 (and eliminating the blanks) gives us the cost of producing different levels of output.

<table>
<thead>
<tr>
<th>Widgets (Q)</th>
<th>1.00</th>
<th>2.00</th>
<th>3.00</th>
<th>4.00</th>
</tr>
</thead>
<tbody>
<tr>
<td>Workers (L)</td>
<td>3.25</td>
<td>4.4</td>
<td>5.2</td>
<td>9</td>
</tr>
<tr>
<td>× Wage Rate per hour</td>
<td>$10</td>
<td>$10</td>
<td>$10</td>
<td>$10</td>
</tr>
<tr>
<td>= Cost</td>
<td>$32.50</td>
<td>$44.00</td>
<td>$52.00</td>
<td>$90.00</td>
</tr>
</tbody>
</table>

This same cost function with which we began (shown in Table 1). Figure 2 shows the graph of the cost function.

![Total Cost](image)

*Figure 2. The Total Cost curve for Widgets. This shows cost increasing at an increasing rate as the firm produces more output.*

Try It

Visit this page in your course online to check your understanding.

Now that we have the basic idea of the cost origins and how they are related to production, let's drill down into the details, by examining average, marginal, fixed, and variable costs.

### Average and Marginal Costs

The cost of producing a firm's output depends on how much labor and capital the firm uses. A list of the costs involved in producing cars will look very different from the costs involved in producing computer software or haircuts or fast-food meals.

We can measure costs in a variety of ways. Each way provides its own insight into costs. Sometimes firms need to look at their cost per unit of output, not just their total cost. There are two ways to measure per unit costs. The most intuitive
way is average cost. **Average cost** is the cost on average of producing a given quantity. We define average cost as total cost divided by the quantity of output produced.

\[
AC = \frac{TC}{Q}
\]

If producing two widgets costs a total of $44, the average cost per widget is

\[
\$44/2 = \$22
\]

per widget. The other way of measuring cost per unit is marginal cost. If average cost is the cost of the average unit of output produced, marginal cost is the cost of each individual unit produced. More formally, **marginal cost** is the cost of producing one more unit (or a few more units) of output. Mathematically, marginal cost is the change in total cost divided by the change in output:

\[
MC = \frac{\Delta TC}{\Delta Q}.
\]

If the cost of the first widget is $32.50 and the cost of two widgets is $44, the marginal cost of the second widget is

\[
\$44 - \$32.50 = \$11.50
\]

We can see the Widget Cost table redrawn below with average and marginal cost added.

<table>
<thead>
<tr>
<th>Q</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Cost</td>
<td>$32.50</td>
<td>$44.00</td>
<td>$52.00</td>
<td>$90.00</td>
</tr>
<tr>
<td>Average Cost</td>
<td>$32.50</td>
<td>$22.00</td>
<td>$17.33</td>
<td>$22.50</td>
</tr>
<tr>
<td>Marginal Cost</td>
<td>$32.50</td>
<td>$11.50</td>
<td>$8.00</td>
<td>$38.00</td>
</tr>
</tbody>
</table>

Note that the marginal cost of the first unit of output is always the same as total cost. Figures 3a and 3b show the graphs of average and marginal cost respectively. The typical shape of each is a U-shape, with average/marginal cost falling at low levels of output and rising at higher levels of output.

**Figure 3. Average and Marginal Cost Curves.** Figure 3a shows the average cost of producing widgets based on the data in Table 6. Figure 3b shows the marginal cost of producing widgets. Both average and marginal cost curves typically are U-shaped.

**Try It**

Visit this page in your course online to check your understanding.
Fixed and Variable Costs

Remember, we explained earlier that fixed inputs are those that cannot be easily adjusted, like a building lease, and variable inputs are those that can be changed easily, like pizza ingredients. We can apply these same terms to costs. **Fixed costs** are the costs of the fixed inputs. Fixed costs do not change regardless of the level of production, at least not in the short term. Whether you produce a lot or a little, the fixed costs are the same. One example is the rent on a factory or a retail space. Once you sign the lease, the rent is the same regardless of how much you produce, at least until the lease runs out.

Fixed costs can take many other forms: for example, the cost of machinery or equipment to produce the product, research and development costs to develop new products, even an expense like advertising to popularize a brand name. The level of fixed costs varies according to the specific line of business: for instance, manufacturing computer chips requires an expensive factory, but a local moving and hauling business can get by with almost no fixed costs at all if it rents trucks by the day when needed.

**Variable costs**, on the other hand, are the costs of the variable inputs; they are incurred in the act of producing—the more you produce, the greater the variable cost. Labor is treated as a variable cost, since producing a greater quantity of a good or service typically requires more workers or more work hours. Variable costs would also include raw materials.

As a concrete example of fixed and variable costs, consider a barber shop called “The Clip Joint.” The data for output and costs are shown in Table 7. The fixed costs of operating the barber shop, including the space and equipment, are $160 per day. The variable costs are the costs of hiring barbers, which in our example is $80 per barber each day. The first two columns of the table show the quantity of haircuts the barbershop can produce as it hires additional barbers. The third column shows the fixed costs, which do not change regardless of the level of production. The fourth column shows the variable costs at each level of output. These are calculated by taking the amount of labor hired and multiplying by the wage. For example, two barbers cost: 2 × $80 = $160. Adding together the fixed costs in the third column and the variable costs in the fourth column produces the total costs in the fifth column. So, for example, with two barbers the total cost is: $160 + $160 = $320.

<table>
<thead>
<tr>
<th>Labor</th>
<th>Quantity</th>
<th>Fixed Cost</th>
<th>Variable Cost</th>
<th>Total Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>16</td>
<td>$160</td>
<td>$80</td>
<td>$240</td>
</tr>
<tr>
<td>2</td>
<td>40</td>
<td>$160</td>
<td>$160</td>
<td>$320</td>
</tr>
<tr>
<td>3</td>
<td>60</td>
<td>$160</td>
<td>$240</td>
<td>$400</td>
</tr>
<tr>
<td>4</td>
<td>72</td>
<td>$160</td>
<td>$320</td>
<td>$480</td>
</tr>
<tr>
<td>5</td>
<td>80</td>
<td>$160</td>
<td>$400</td>
<td>$560</td>
</tr>
<tr>
<td>6</td>
<td>84</td>
<td>$160</td>
<td>$480</td>
<td>$640</td>
</tr>
<tr>
<td>7</td>
<td>82</td>
<td>$160</td>
<td>$560</td>
<td>$720</td>
</tr>
</tbody>
</table>

The relationship between the quantity of output being produced and the cost of producing that output is shown graphically in the Figure 4. The fixed costs are always shown as the vertical intercept of the total cost curve; that is, they are the costs incurred when output is zero so there are no variable costs. You can see from the graph that once production starts, total costs and variable costs rise. While variable costs may initially increase at a decreasing rate, at some point they begin increasing at an increasing rate. This is caused by diminishing marginal returns, which is easiest to see with an example. As the number of barbers increases from zero to one in the table, output increases from 0 to 16 for a marginal gain of 16; as the number rises from one to two barbers, output increases from 16 to 40, a marginal gain of 24. From that point on, though, the marginal gain in output diminishes as each additional barber is added.
Factor payments:

- **Fixed cost**: Cost of the fixed inputs; expenditure that a firm must make before production starts and that does not change regardless of the production level.

- **Marginal cost**: The additional cost of producing one more unit; mathematically, \( MC = \Delta TC / \Delta Q \).

---

**Figure 4. “The Clip Joint” Total Costs.** At zero production, the fixed costs of $160 are still present. As production increases, variable costs are added to fixed costs, and the total cost is the sum of the two.

For example, as the number of barbers rises from two to three, the marginal output gain is only 20; and as the number rises from three to four, the marginal gain is only 12. To understand the reason behind this pattern, consider that a one-man barber shop is a very busy operation. The single barber needs to do everything: say hello to people entering, answer the phone, cut hair, sweep up, and run the cash register. A second barber reduces the level of disruption from jumping back and forth between these tasks, and allows a greater division of labor and specialization. The result can be greater increasing marginal returns. However, as other barbers are added, the advantage of each additional barber is less, since the specialization of labor can only go so far. The addition of a sixth or seventh or eighth barber just to greet people at the door will have less impact than the second one did. This is the pattern of diminishing marginal returns. At some point, you may even see negative returns as the additional barbers begin bumping elbows and getting in each other’s way. In this case, the addition of still more barbers would actually cause output to decrease, as shown in the last row of Table 1. As a result, the total costs of production will begin to rise more rapidly as output increases.

This pattern of **diminishing marginal returns** is common in production. As another example, consider the problem of irrigating a crop on a farmer’s field. The plot of land is the fixed factor of production, while the water that can be added to the land is the key variable cost. As the farmer adds water to the land, output increases. But adding more and more water brings smaller and smaller increases in output, until at some point the water floods the field and actually reduces output. Diminishing marginal returns occur because, at a given level of fixed costs, each additional input contributes less and less to overall production.

---

**Try It**

Visit this page in your course online to check your understanding.

**Watch It**

Watch this clip to review and assess your ability to identify the variable, fixed, total, and marginal costs.

Visit this page in your course online to view this presentation.

**Glossary**

- **Factor payments**: What the firm pays for the use of the factors of production—includes raw materials, rent, wages and salaries, interest and dividends, and profit for entrepreneurship.

- **Fixed cost**: Cost of the fixed inputs; expenditure that a firm must make before production starts and that does not change regardless of the production level.

- **Marginal cost**: The additional cost of producing one more unit; mathematically, \( MC = \Delta TC / \Delta Q \).
**total cost**: the sum of fixed and variable costs of production

**variable cost**: cost of production that increases with the quantity produced; the cost of the variable inputs

---

**LEARN BY DOING: COSTS IN THE SHORT RUN**

**Try It**

These questions allow you to get as much practice as you need, as you can click the link at the top of the first question (“Try another version of these questions”) to get a new set of questions. Practice until you feel comfortable doing the questions.

Visit this page in your course online to practice before taking the quiz.

---

**AVERAGE COSTS AND CURVES**

**Learning Objectives**

- Describe and calculate average total costs and average variable costs
- Calculate and graph marginal cost
- Analyze the relationship between marginal and average costs

The cost of producing a firm’s output depends on how much labor and capital the firm uses. A list of the costs involved in producing cars will look very different from the costs involved in producing computer software or haircuts or fast-food meals. However, the cost structure of all firms can be broken down into some common underlying patterns. When a firm
looks at its total costs of production in the short run, a useful starting point is to divide total costs into two categories: fixed costs that cannot be changed in the short run and variable costs that can be changed.

The breakdown of total costs into fixed and variable costs can provide a basis for other insights as well. The first five columns of Table 1 should look familiar—they come from the Clip Joint example we saw earlier—but there are also three new columns showing average total costs, average variable costs, and marginal costs. These new measures analyze costs on a per-unit (rather than a total) basis.

<table>
<thead>
<tr>
<th>Labor</th>
<th>Quantity</th>
<th>Fixed Cost</th>
<th>Variable Cost</th>
<th>Total Cost</th>
<th>Marginal Cost</th>
<th>Average Total Cost</th>
<th>Average Variable Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>16</td>
<td>$160</td>
<td>$80</td>
<td>$240</td>
<td>$5.00</td>
<td>$15.00</td>
<td>$5.00</td>
</tr>
<tr>
<td>2</td>
<td>40</td>
<td>$160</td>
<td>$160</td>
<td>$320</td>
<td>$3.30</td>
<td>$8.00</td>
<td>$4.00</td>
</tr>
<tr>
<td>3</td>
<td>60</td>
<td>$160</td>
<td>$240</td>
<td>$400</td>
<td>$4.00</td>
<td>$6.60</td>
<td>$4.00</td>
</tr>
<tr>
<td>4</td>
<td>72</td>
<td>$160</td>
<td>$320</td>
<td>$480</td>
<td>$6.60</td>
<td>$6.60</td>
<td>$4.40</td>
</tr>
<tr>
<td>5</td>
<td>80</td>
<td>$160</td>
<td>$400</td>
<td>$560</td>
<td>$10.00</td>
<td>$7.00</td>
<td>$5.00</td>
</tr>
<tr>
<td>6</td>
<td>84</td>
<td>$160</td>
<td>$480</td>
<td>$640</td>
<td>$20.00</td>
<td>$7.60</td>
<td>$5.70</td>
</tr>
</tbody>
</table>

**Watch It**

Watch this clip as a continuation from the video on the previous page to see how average variable cost, average fixed costs, and average total costs are calculated.

Visit this page in your course online to view this presentation.

**Average total cost** is total cost divided by the quantity of output. Since the total cost of producing 40 haircuts at “The Clip Joint” is $320, the average total cost for producing each of 40 haircuts is $320/40, or $8 per haircut. Average cost curves are typically U-shaped, as Figure 1 shows. Average total cost starts off relatively high, because at low levels of output total costs are dominated by the fixed cost; mathematically, the denominator is so small that average total cost is large. Average total cost then declines, as the fixed costs are spread over an increasing quantity of output. In the average cost calculation, the rise in the numerator of total costs is relatively small compared to the rise in the denominator of quantity produced. But as output expands still further, the average cost begins to rise. At the right side of the average cost curve, total costs begin rising more rapidly as diminishing returns kick in.
Figure 1. Cost Curves at the Clip Joint. The information on total costs, fixed cost, and variable cost can also be presented on a per-unit basis. Average total cost (ATC) is calculated by dividing total cost by the total quantity produced. The average total cost curve is typically U-shaped. Average variable cost (AVC) is calculated by dividing variable cost by the quantity produced. The average variable cost curve lies below the average total cost curve and is typically U-shaped or upward-sloping. Marginal cost (MC) is calculated by taking the change in total cost between two levels of output and dividing by the change in output. The marginal cost curve is upward-sloping.

Average variable cost obtained when variable cost is divided by quantity of output. For example, the variable cost of producing 80 haircuts is $400, so the average variable cost is $400/80, or $5 per haircut. Note that at any level of output, the average variable cost curve will always lie below the curve for average total cost, as shown in Figure 1. The reason is that average total cost includes average variable cost and average fixed cost. Thus, for Q = 80 haircuts, the average total cost is $8 per haircut, while the average variable cost is $5 per haircut. However, as output grows, fixed costs become relatively less important (since they do not rise with output), so average variable cost sneaks closer to average cost. Average total and variable costs measure the average costs of producing some quantity of output. Marginal cost is somewhat different.

Try It

Visit this page in your course online to check your understanding.

Recall that marginal cost, which we introduced on the previous page, is the additional cost of producing one more unit of output. So it is not the cost per unit of all units being produced, but only the next one (or next few). Marginal cost can be calculated by taking the change in total cost and dividing it by the change in quantity. For example, as quantity produced increases from 40 to 60 haircuts, total costs rise by 400 – 320, or 80. Thus, the marginal cost for each of those marginal 20 units will be 80/20, or $4 per haircut.

The marginal cost curve may fall for the first few units of output but after that are generally upward-sloping, because diminishing marginal returns implies that additional units are more costly to produce. A small range of increasing marginal returns can be seen in the figure as a dip in the marginal cost curve before it starts rising.

Watch It

Watch this video to learn how to draw the various cost curves, including total, fixed and variable costs, marginal cost, average total, average variable, and average fixed costs.

Watch this video online: https://youtu.be/qYKJdooEnwU
Where do marginal and average costs meet?

The marginal cost curve intersects the average cost curve exactly at the bottom of the average cost curve—which occurs at a quantity of 72 and cost of $6.60 in Figure 1. The reason why the intersection occurs at this point is built into the economic meaning of marginal and average costs. If the marginal cost of production is below the average cost for producing previous units, as it is for the points to the left of where MC crosses ATC, then producing one more additional unit will reduce average costs overall—and the ATC curve will be downward-sloping in this zone. Conversely, if the marginal cost of production for producing an additional unit is above the average cost for producing the earlier units, as it is for points to the right of where MC crosses ATC, then producing a marginal unit will increase average costs overall—and the ATC curve must be upward-sloping in this zone. The point of transition, between where MC is pulling ATC down and where it is pulling it up, must occur at the minimum point of the ATC curve.

The same relationship is true for marginal cost and average variable cost. The reasoning is the same also. This does not hold for average fixed cost. Do you know why not? It’s because marginal cost affects variable cost, but it does not affect fixed cost.

This idea of the marginal cost “pulling down” the average cost or “pulling up” the average cost may sound abstract, but think about it in terms of your own grades. If the score on the most recent quiz you take is lower than your average score on previous quizzes, then the marginal quiz pulls down your average. If your score on the most recent quiz is higher than the average on previous quizzes, the marginal quiz pulls up your average. In this same way, low marginal costs of production first pull down average costs and then higher marginal costs pull them up.

The numerical calculations behind average cost, average variable cost, and marginal cost will change from firm to firm. However, the general patterns of these curves, and the relationships and economic intuition behind them, will not change.

Why are total cost and average cost not on the same graph?

Total cost, fixed cost, and variable cost each reflect different aspects of the cost of production over the entire quantity of output being produced. These costs are measured in dollars. In contrast, marginal cost, average cost, and average variable cost are costs per unit. In the previous example, they are measured as cost per haircut. Thus, it would not make sense to put all of these numbers on the same graph, since they are measured in different units ($ versus $ per unit of output).

It would be as if the vertical axis measured two different things. In addition, as a practical matter, if they were on the same graph, the lines for marginal cost, average cost, and average variable cost would appear almost flat against the horizontal axis, compared to the values for total cost, fixed cost, and variable cost. Using the figures from the previous example, the total cost of producing 40 haircuts is $320. But the average cost is $320/40, or $8. If you graphed both total and average cost on the same axes, the average cost would hardly show.

Glossary

- **average total cost**: for any quantity of output, total cost divided by the quantity of output
- **average variable cost**: for any quantity of output, variable cost divided by the quantity of output
LESSONS FROM ALTERNATIVE MEASURES OF COST

Learning Objectives

- Explain profit margin

Lessons from Alternative Measures of Cost

Breaking down total costs into fixed cost, marginal cost, average total cost, and average variable cost is useful because each statistic offers its own insights for the firm. Whatever the firm’s quantity of production, total revenue must exceed total costs if it is to earn a profit.

Fixed costs are often **sunk costs** that, once incurred, cannot be recouped. In thinking about what to do next, sunk costs should typically be ignored, since this spending has already been made and cannot be changed. However, variable costs can be changed, so they convey information about the firm’s ability to cut costs in the present and the extent to which costs will increase if production rises.

Profit Margin

Businesses often talk about their profit margin (or average profit). Profit margin tells a firm for any level of output are they making money or losing money. Profit is defined as revenues minus costs.

\[
total \ profit = total \ revenue - total \ cost
\]
Finding Average Profit

Starting with the equation for total profit above, if we divide both sides of the total profit equation by quantity, we get average profit:

\[
\text{average profit} = \frac{\text{total profit}}{\text{quantity produced}} = \frac{\text{total revenue} - \text{total cost}}{\text{quantity produced}}
\]

If we separate the two parts of the right hand side we get:

\[
\text{average profit} = \frac{\text{total revenue}}{\text{quantity produced}} - \frac{\text{total cost}}{\text{quantity produced}}
\]

Or, in other words:

\[
\text{average profit} = \text{average revenue} - \text{average cost}
\]

But, average revenue is just another name for price, as shown below:

\[
\text{average revenue} = \frac{\text{price} \times \text{quantity produced}}{\text{quantity produced}} = \text{price}
\]

Substituting this back into the equation for average profit above, we get:

\[
\text{average profit} = \text{price} - \text{average cost}
\]

Average profit is the firm’s profit margin. What this means is that if, at a given level of output, the market price is above average cost, average profit and thus total profit, will be positive; in other words, the firm is making money. We’ll be able to show this graphically beginning in the next module, but for now, let’s look at it numerically. If, for a given level of output, price is below average cost, then profits will be negative, and the firm is losing money.

Suppose a firm is producing 100 units of output at a price of $5 each. Suppose the firm’s average cost is $3 per unit of output. Since the $5 price is greater than the $3 average cost, we can immediately tell that the firm’s profit margin is $2 for each of the 100 units of output produced and sold. Thus, the firm’s total profit is $2 per unit times 100 units = $200.

Alternatively, we could compute the total revenue as price times quantity = $5 per unit x 100 units = $500. Similarly, we could compute the total cost as average cost times quantity = $3 per unit x 100 units = $300. Thus, the firm’s total profit is total revenue minus total cost = $500 minus $300 = $200, which you’ll note is the same answer we got by using the profit margin.

Try It

Visit this page in your course online to check your understanding.

A Variety of Cost Patterns

The pattern of costs varies among industries and even among firms in the same industry. Some businesses have high fixed costs, but low marginal costs. Consider, for example, an Internet company that provides medical advice to customers. Such a company might be paid by consumers directly, or perhaps hospitals or healthcare practices might subscribe on behalf of their patients.

Setting up the website, collecting the information, writing the content, and buying or leasing the computer space to handle the web traffic are all fixed costs that must be undertaken before the site can work. However, when the website is up and running, it can provide a high quantity of service with relatively low variable costs, like the cost of monitoring the system and updating the information.

In this case, the total cost curve might start at a high level, because of the high fixed costs, but then might appear close to flat, up to a large quantity of output, reflecting the low variable costs of operation. If the website is popular, however, a large rise in the number of visitors will overwhelm the website, and increasing output further could require a purchase of additional computer space.
For other firms, fixed costs may be relatively low. For example, consider firms that rake leaves in the fall or shovel snow off sidewalks and driveways in the winter. For fixed costs, such firms may need little more than a car to transport workers to homes of customers and some rakes and shovels. Still other firms may find that diminishing marginal returns set in quite sharply. If a manufacturing plant tried to run 24 hours a day, seven days a week, little time remains for routine maintenance of the equipment, and marginal costs can increase dramatically as the firm struggles to repair and replace overworked equipment.

Every firm can gain insight into its task of earning profits by dividing its total costs into fixed and variable costs, and then using these calculations as a basis for average total cost, average variable cost, and marginal cost. However, making a final decision about the profit-maximizing quantity to produce and the price to charge will require combining these perspectives on cost with an analysis of sales and revenue, which in turn requires looking at the market structure in which the firm finds itself. Before we turn to the analysis of market structure in other modules, we will analyze the firm’s cost structure from a long-run perspective.

**gLossary**

**profit margin**: at a given level of output, the difference between price and average cost; also known as average profit

**sunk costs**: fixed costs that, once incurred, cannot be recouped

**INTRODUCTION TO PRODUCTION AND COSTS IN THE LONG RUN**

What you’ll learn to do: examine production choices in the long run

In the long run, because there are no fixed inputs, firms have more options about how to produce their products. As a result, we will see that long run costs are typically less than short run costs.
PRODUCTION IN THE LONG RUN

Learning Objectives

- Describe how long run production differs from short run production

In the long run, all factors (including capital) are variable, so our production function is \( Q = f[L,K] \).

Consider a secretarial firm that does typing for hire using typists for labor and personal computers for capital. To start, the firm has just enough business for one typist and one PC to keep busy for a day. Say that’s five documents. Now suppose the firm receives a rush order from a good customer for 10 documents tomorrow. Ideally, the firm would like to use two typists and two PCs to produce twice their normal output of five documents. However, in the short turn, the firm has fixed capital, i.e. only one PC. The table below shows the situation:

<table>
<thead>
<tr>
<th>Labor (# Typists)</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Product (Letters/hr)</td>
<td>5</td>
<td>7</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>For K = 1 PC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Marginal Product</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

In the short run, the only variable factor is labor so the only way the firm can produce more output is by hiring additional workers. What could the second worker do? What can they contribute to the firm? Perhaps they can answer the phone, which is a major impediment to completing the typing assignment. What about a third worker? Perhaps he or she could bring coffee to the first two workers. You can see both total product and marginal product for the firm above. Now here’s something to think about: at what point (e.g. after how many workers) does diminishing marginal productivity kick in, and more importantly, why?
In this example, marginal productivity starts to decline after the second worker. This is because capital is fixed at one PC. The production process for typing works best when each worker is combined with one PC. If you add more than one typist, you get seriously diminishing marginal productivity.

Now consider the long run. Suppose the firm’s demand increases to 15 documents per day. What might the firm do to operate more efficiently? If demand has tripled, the firm could acquire two more PCs, which would give us a new short run production function as Table 2 below shows.

<table>
<thead>
<tr>
<th>Labor (# Typists)</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Product (Letters/hr)</td>
<td>5</td>
<td>6</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>Marginal Product</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total Product (Letters/hr)</td>
<td>5</td>
<td>10</td>
<td>15</td>
<td>17</td>
<td>18</td>
<td>18</td>
</tr>
<tr>
<td>Marginal Product</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

For K = 1 PC

For K = 3 PC

With more capital, the firm can hire three workers before diminishing productivity comes into effect. More generally, because all factors are variable, the long run production function shows the most efficient way of producing any level of output.

Try It

Visit this page in your course online to check your understanding.

LONG RUN COSTS AND PRODUCTION TECHNOLOGY
Learning Objectives

- Explain the impact of production technology on long run total costs

Long Run Costs

The long run is the period of time when all costs are variable. The long run depends on the specifics of the firm in question—it is not a precise period of time. If you have a one-year lease on your factory, then the long run is any period longer than a year, since after a year you are no longer bound by the lease. No costs are fixed in the long run. A firm can build new factories and purchase new machinery, or it can close existing facilities. In planning for the long run, the firm will compare alternative production technologies (or processes).

In this context, technology refers to all alternative methods of combining inputs to produce outputs. It does not refer to a specific new invention like the tablet computer. The firm will search for the production technology that allows it to produce the desired level of output at the lowest cost. After all, lower costs lead to higher profits—at least if total revenues remain unchanged. Moreover, each firm must fear that if it does not seek out the lowest-cost methods of production, then it may lose sales to competitor firms that find a way to produce and sell for less.

Choice of Production Technology

Many tasks can be performed with a range of combinations of labor and physical capital. For example, a firm can have human beings answering phones and taking messages, or it can invest in an automated voicemail system. A firm can hire file clerks and secretaries to manage a system of paper folders and file cabinets, or it can invest in a computerized record keeping system that will require fewer employees. A firm can hire workers to push supplies around a factory on rolling carts, it can invest in motorized vehicles, or it can invest in robots that carry materials without a driver. Firms often face a choice between buying many small machines, which need a worker to run each one, or buying one larger and more expensive machine, which requires only one or two workers to operate it. In short, physical capital and labor can often substitute for each other.

Consider the example of a private firm that is hired by local governments to clean up public parks. Three different combinations of labor and physical capital for cleaning up a single average-sized park appear in Table 1. The first production technology is heavy on workers and light on machines, while the next two technologies substitute machines for workers. Since all three of these production methods produce the same thing—one cleaned-up park—a profit-seeking firm will choose the production technology that is least expensive, given the prices of labor and machines.

<table>
<thead>
<tr>
<th>Production technology 1</th>
<th>10 workers</th>
<th>2 machines</th>
</tr>
</thead>
<tbody>
<tr>
<td>Production technology 2</td>
<td>7 workers</td>
<td>4 machines</td>
</tr>
<tr>
<td>Production technology 3</td>
<td>3 workers</td>
<td>7 machines</td>
</tr>
</tbody>
</table>

Production technology 1 uses the most labor and least machinery, while production technology 3 uses the least labor and the most machinery. Table 2 outlines three examples of how the total cost will change with each production technology as the cost of labor changes. As the cost of labor rises from example A to B to C, the firm will choose to substitute away from labor and use more machinery.

<table>
<thead>
<tr>
<th>Example A: Workers cost $40, machines cost $80</th>
</tr>
</thead>
<tbody>
<tr>
<td>Labor Cost</td>
</tr>
<tr>
<td>------------</td>
</tr>
<tr>
<td>Cost of technology 1</td>
</tr>
</tbody>
</table>
### Example A: Workers cost $40, machines cost $80

<table>
<thead>
<tr>
<th>Cost of technology 1</th>
<th>Labor Cost</th>
<th>Machine Cost</th>
<th>Total Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 × $55 = $550</td>
<td>2 × $80 = $160</td>
<td></td>
<td>$710</td>
</tr>
<tr>
<td>Cost of technology 2</td>
<td>7 × $55 = $385</td>
<td>4 × $80 = $320</td>
<td>$705</td>
</tr>
<tr>
<td>Cost of technology 3</td>
<td>3 × $55 = $165</td>
<td>7 × $80 = $560</td>
<td>$725</td>
</tr>
</tbody>
</table>

### Example B: Workers cost $55, machines cost $80

<table>
<thead>
<tr>
<th>Cost of technology 1</th>
<th>Labor Cost</th>
<th>Machine Cost</th>
<th>Total Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 × $55 = $550</td>
<td>2 × $80 = $160</td>
<td></td>
<td>$710</td>
</tr>
<tr>
<td>Cost of technology 2</td>
<td>7 × $55 = $385</td>
<td>4 × $80 = $320</td>
<td>$705</td>
</tr>
<tr>
<td>Cost of technology 3</td>
<td>3 × $55 = $165</td>
<td>7 × $80 = $560</td>
<td>$725</td>
</tr>
</tbody>
</table>

### Example C: Workers cost $90, machines cost $80

<table>
<thead>
<tr>
<th>Cost of technology 1</th>
<th>Labor Cost</th>
<th>Machine Cost</th>
<th>Total Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 × $90 = $900</td>
<td>2 × $80 = $160</td>
<td></td>
<td>$1,060</td>
</tr>
<tr>
<td>Cost of technology 2</td>
<td>7 × $90 = $630</td>
<td>4 × $80 = $320</td>
<td>$950</td>
</tr>
<tr>
<td>Cost of technology 3</td>
<td>3 × $90 = $270</td>
<td>7 × $80 = $560</td>
<td>$830</td>
</tr>
</tbody>
</table>

Example A shows the firm’s cost calculation when wages are $40 and machines costs are $80. In this case, technology 1 is the low-cost production technology. In example B, wages rise to $55, while the cost of machines does not change, in which case technology 2 is the low-cost production technology. If wages keep rising up to $90, while the cost of machines remains unchanged, then technology 3 clearly becomes the low-cost form of production, as example C shows.

This example shows that as an input becomes more expensive (in this case, the labor input), firms will attempt to conserve on using that input and will instead shift to other inputs that are relatively less expensive. This pattern helps to explain why the demand curve for labor (or any input) slopes down; that is, as labor becomes relatively more expensive, profit-seeking firms will seek to substitute the use of other inputs. When a multinational employer like Coca-Cola or McDonald’s sets up a bottling plant or a restaurant in a high-wage economy like the United States, Canada, Japan, or Western Europe, it is likely to use production technologies that conserve on the number of workers and focuses more on machines. However, that same employer is likely to use production technologies with more workers and less machinery when producing in a lower-wage country like Mexico, China, or South Africa.

One final point: an **improvement** in production technology is a new method of production, or a new process, that produces more output with the same amount of inputs, or it produces the same output using less inputs. Thus, an improvement in production technology leads to a reduction in production cost.

---

### Glossary

**production technologies**: alternative methods of combining inputs to produce output

---

**Licensing & Attributions**

CC licensed content, Shared previously

- Costs in the Long Run. **Authored by**: OpenStax College. **Located at**: https://cnx.org/contents/vEmOH_...p@4.44:9SU8XZql@9/Costs-in-the-Long-Run. **License**: CC BY: Attribution
ECONOMIES OF SCALE

Learning Objectives

- Identify economies of scale, diseconomies of scale, and constant returns to scale

Economies of Scale

Earlier in this module we saw that in the short run when a firm increases its scale of operation (or its level of output), its average cost of production can decrease or increase. This is illustrated in Figure 1.

![Figure 1. Short Run Average Costs. The normal shape for a short-run average cost curve is U-shaped with decreasing average costs at low levels of output and increasing average costs at high levels of output.](image)

What happens to a firm’s average costs when it increases its level of output in the long run? Many industries experience economies of scale. **Economies of scale** refers to the situation where, as the quantity of output goes up, the cost per unit goes down. This is the idea behind “warehouse stores” like Costco or Walmart. In everyday language: a larger factory can produce at a lower average cost than a smaller factory. Figure 1 illustrates the idea of economies of scale, showing the average cost of producing an alarm clock falling as the quantity of output rises. For a small-sized factory like S, with an output level of 1,000, the average cost of production is $12 per alarm clock. For a medium-sized factory like M, with an output level of 2,000, the average cost of production falls to $8 per alarm clock. For a large factory like L, with an output of 5,000, the average cost of production declines still further to $4 per alarm clock.
Figure 2. Economies of Scale  A small factory like S produces 1,000 alarm clocks at an average
cost of $12 per clock. A medium factory like M produces 2,000 alarm clocks at a cost of $8 per clock.
A large factory like L produces 5,000 alarm clocks at a cost of $4 per clock. Economies of scale exist
because the larger scale of production leads to lower average costs.

The average cost curve in Figure 2 may appear similar to the average cost curve in Figure 1, although it is downward-sloping rather than U-shaped. But there is one major difference. The economies of scale curve is a long-run average cost curve, because it allows all factors of production to change. Short-run average cost curves assume the existence of fixed costs, and only variable costs were allowed to change. In sum, economies of scale refers to a situation where long run average cost decreases as the firm’s output increases.

One prominent example of economies of scale occurs in the chemical industry. Chemical plants have a lot of pipes. The cost of the materials for producing a pipe is related to the circumference of the pipe and its length. However, the volume of chemicals that can flow through a pipe is determined by the cross-section area of the pipe. The calculations in Table 1 show that a pipe which uses twice as much material to make (as shown by the circumference of the pipe doubling) can actually carry four times the volume of chemicals because the cross-section area of the pipe rises by a factor of four (as shown in the Area column).

<table>
<thead>
<tr>
<th>Circumference (2πr)</th>
<th>Area (πr²)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4-inch pipe</td>
<td>12.5 inches</td>
</tr>
<tr>
<td>8-inch pipe</td>
<td>25.1 inches</td>
</tr>
<tr>
<td>16-inch pipe</td>
<td>50.2 inches</td>
</tr>
</tbody>
</table>

A doubling of the cost of producing the pipe allows the chemical firm to process four times as much material. This pattern is a major reason for economies of scale in chemical production, which uses a large quantity of pipes. Of course, economies of scale in a chemical plant are more complex than this simple calculation suggests. But the chemical engineers who design these plants have long used what they call the “six-tenths rule,” a rule of thumb which holds that increasing the quantity produced in a chemical plant by a certain percentage will increase total cost by only six-tenths as much.

Watch It
Watch this video to see an example of economies of scale as applied to making bread.
Watch this video online: https://youtu.be/JdCgu1sOPDo

Shapes of Long-Run Average Cost Curves
While in the short run firms are limited to operating on a single average cost curve (corresponding to the level of fixed costs they have chosen), in the long run when all costs are variable, they can choose to operate on any average cost curve. Thus, the long-run average cost (LRAC) curve is actually based on a group of short-run average cost (SRAC) curves, each of which represents one specific level of fixed costs. More precisely, the long-run average cost curve will be the least expensive average cost curve for any level of output. Figure 3 shows how the long-run average cost curve is built from a group of short-run average cost curves.

Five short-run-average cost curves appear on the diagram. Each SRAC curve represents a different level of fixed costs. For example, you can imagine SRAC\(_1\) as a small factory, SRAC\(_2\) as a medium factory, SRAC\(_3\) as a large factory, and SRAC\(_4\) and SRAC\(_5\) as very large and ultra-large. Although this diagram shows only five SRAC curves, presumably there are an infinite number of other SRAC curves between the ones that we show. Think of this family of short-run average cost curves as representing different choices for a firm that is planning its level of investment in fixed cost—knowing that different choices about capital investment in the present will cause it to end up with different short-run average cost curves in the future.

![Figure 3. From Short-Run Average Cost Curves to Long-Run Average Cost Curves](image)

The five different short-run average cost (SRAC) curves each represents a different level of fixed costs, from the low level of fixed costs at SRAC\(_1\) to the high level of fixed costs at SRAC\(_5\). Other SRAC curves, not shown in the diagram, lie between the ones that are shown here. The long-run average cost (LRAC) curve shows the lowest cost for producing each quantity of output when fixed costs can vary, and so it is formed by the bottom edge of the family of SRAC curves. If a firm wished to produce quantity Q\(_3\), it would choose the fixed costs associated with SRAC\(_3\).

The long-run average cost curve shows the cost of producing each quantity in the long run, when the firm can choose its level of fixed costs and thus choose which short-run average costs it desires. If the firm plans to produce in the long run at an output of Q\(_3\), it should make the set of investments that will lead it to locate on SRAC\(_3\), which allows producing q\(_3\) at the lowest cost. A firm that intends to produce Q\(_3\) would be foolish to choose the level of fixed costs at SRAC\(_2\) or SRAC\(_4\). At SRAC\(_2\) the level of fixed costs is too low for producing Q\(_3\) at lowest possible cost, and producing q\(_3\) would require adding a very high level of variable costs and make the average cost very high. At SRAC\(_4\), the level of fixed costs is too high for producing q\(_3\) at lowest possible cost, and again average costs would be very high as a result.

The shape of the long-run cost curve, in Figure 3, is fairly common for many industries. The left-hand portion of the long-run average cost curve, where it is downward-sloping from output levels Q\(_1\) to Q\(_2\) to Q\(_3\), illustrates the case of economies of scale. In this portion of the long-run average cost curve, larger scale leads to lower average costs. We illustrated this pattern earlier in Figure 2.

In the middle portion of the long-run average cost curve, the flat portion of the curve around Q\(_3\), economies of scale have been exhausted. In this situation, allowing all inputs to expand does not much change the average cost of production. We call this constant returns to scale. In this LRAC curve range, the average cost of production does not change much as scale rises or falls.
How do Economies of Scale Compare to Diminishing Marginal Returns?

The concept of economies of scale, where average costs decline as production expands, might seem to conflict with the idea of diminishing marginal returns, where marginal costs rise as production expands. But diminishing marginal returns refers only to the short-run average cost curve, where one variable input (like labor) is increasing, but other inputs (like capital) are fixed. Economies of scale refers to the long-run average cost curve where all inputs are being allowed to increase together. Thus, it is quite possible and common to have an industry that has both diminishing marginal returns when only one input is allowed to change, and at the same time has increasing or constant economies of scale when all inputs change together to produce a larger-scale operation.

Finally, the right-hand portion of the long-run average cost curve, running from output level Q4 to Q5, shows a situation where, as the level of output and the scale rises, average costs rise as well. This situation is called **diseconomies of scale**. A firm or a factory can grow so large that it becomes very difficult to manage, resulting in unnecessarily high costs as many layers of management try to communicate with workers and with each other, and as failures to communicate lead to disruptions in the flow of work and materials. Not many overly large factories exist in the real world, because with their very high production costs, they are unable to compete for long against plants with lower average costs of production. However, in some planned economies, like the economy of the old Soviet Union, plants that were so large as to be grossly inefficient were able to continue operating for a long time because government economic planners protected them from competition and ensured that they would not make losses.

Diseconomies of scale can also be present across an entire firm, not just a large factory. The **leviathan effect** can hit firms that become too large to run efficiently, across the entirety of the enterprise. Firms that shrink their operations are often responding to finding itself in the diseconomies region, thus moving back to a lower average cost at a lower output level.

**Try It**

Visit this page in your course online to check your understanding.

**Glossary**

- **constant returns to scale**: expanding all inputs proportionately does not change the average cost of production
- **economies of scale**: the long-run average cost of producing output decreases as total output increases
- **diseconomies of scale**: the long-run average cost of producing output increases as total output increases
- **leviathan effect**: when a firm gets so large that it operates inefficiently, experiencing diseconomies of scale
- **long-run average cost (LRAC) curve**: shows the lowest possible average cost of production, allowing all the inputs to production to vary so that the firm is choosing its production technology
- **short-run average cost (SRAC) curve**: the average total cost curve in the short term; shows the total of the average fixed costs and the average variable costs
Learning Objectives

- Describe how the shape of the long-run average cost curve affects the number of firms that an industry can sustain and the market structure in the industry

The Size and Number of Firms in an Industry

The shape of the long-run average cost curve has implications for how many firms will compete in an industry, and whether the firms in an industry have many different sizes, or tend to be the same size. For example, say that one million dishwashers are sold every year at an average cost of $500 each and the long-run average cost curve for dishwashers is shown in Figure 1(a). In Figure 1(a), the lowest point of a firm's LRAC curve occurs at a quantity of 10,000 produced. Thus, the market for dishwashers will consist of 100 different manufacturing plants of this same size. If some firms built a plant that produced 5,000 dishwashers per year or 25,000 dishwashers per year, Figure 1(a) shows that the average cost of production at such plants would be well above $500, and the firms would not be able to compete with firms that produced 10,000 dishwashers at an average cost of $500.

![Figure 1. The LRAC Curve and the Size and Number of Firms.](image)

- (a) LRAC curve with a clear minimum point. When the LRAC curve has a clear minimum point, then any firm producing a different quantity will have higher costs. In this case, a firm producing at a quantity of 10,000 will produce at a lower average cost than a firm producing, say, 5,000 or 20,000 units. (b) Low-cost firms will produce between output levels R and S. When the LRAC curve has a flat bottom, then firms producing at any quantity along this flat bottom can compete. In this case, any firm producing a quantity between 5,000 and 20,000 can compete effectively, although firms producing less than 5,000 or more than 20,000 would face higher average costs and be unable to compete.

A more common case is illustrated in Figure 1(b), where the LRAC curve has a flat-bottomed area of constant returns to scale. In this situation, any firm with a level of output between 5,000 and 20,000 will be able to produce at about the same level of average cost. Given that the market demand for dishwashers is one million per year, this market might have as many as 200 producers (that is, one million dishwashers divided by firms making 5,000 each) or as few as 50 producers (one million dishwashers divided by firms making 20,000 each). The producers in this market will range in size from firms that make 5,000 units to firms that make 20,000 units. But firms that produce below 5,000 units or more than 20,000 will be unable to compete, because their average costs will be too high. Thus, if we see an industry where almost all plants are the same size, it is likely that the long-run average cost curve has a unique bottom point as in Figure 1(a). However, if the long-run average cost curve has a wide flat bottom like Figure 1(b), then firms of a variety of different sizes will be able to compete with each other.
The flat section of the long-run average cost curve in Figure 1(b) can be interpreted in two different ways. One interpretation is that a single manufacturing plant producing a quantity of 5,000 has the same average costs as a single manufacturing plant with four times as much capacity that produces a quantity of 20,000. The other interpretation is that one firm owns a single manufacturing plant that produces a quantity of 5,000, while another firm owns four separate manufacturing plants, which each produce a quantity of 5,000. This second explanation, based on the insight that a single firm may own a number of different manufacturing plants, is especially useful in explaining why the long-run average cost curve often has a large flat segment—and thus why a seemingly smaller firm may be able to compete quite well with a larger firm. At some point, however, the task of coordinating and managing many different plants raises the cost of production sharply, and the long-run average cost curve slopes up as a result.

In the examples to this point, the quantity demanded in the market is quite large (one million) compared with the quantity produced at the bottom of the long-run average cost curve (5,000, 10,000 or 20,000). In such a situation, the market is set for competition between many firms. But what if the bottom of the long-run average cost curve is at a quantity of 10,000 and the total market demand at that price is only slightly higher than that quantity—or even somewhat lower?

Return to Figure 1(a), where the bottom of the long-run average cost curve is at 10,000, but now imagine that the total quantity of dishwashers demanded in the market is only 30,000. In this situation, the total number of firms in the market would be three.

Alternatively, consider a situation, again in the setting of Figure 1(a), where the bottom of the long-run average cost curve is 10,000, but total demand for the product is only 5,000. (For simplicity, imagine that this demand is highly inelastic, so that it does not vary according to price.) In this situation, the market may well end up with a single firm—a monopoly—producing all 5,000 units. If any firm tried to challenge this monopoly while producing a quantity lower than 5,000 units, the prospective competitor firm would have a higher average cost, and so it would not be able to compete in the longer term without losing money.

Thus, the shape of the long-run average cost curve reveals whether competitors in the market will be different sizes. If the LRAC curve has a single point at the bottom, then the firms in the market will be about the same size, but if the LRAC curve has a flat-bottomed segment of constant returns to scale, then firms in the market may be a variety of different sizes. The relationship between the quantity at the minimum of the long-run average cost curve and the quantity demanded in the market will predict how much competition is likely to exist in the market. If the quantity demanded in the market far exceeds the quantity at the minimum of the LRAC, then many firms will compete. If the quantity demanded in the market is only slightly higher than the quantity at the minimum of the LRAC, a few firms will compete. If the quantity demanded in the market is less than the quantity at the minimum of the LRAC, a single-producer monopoly is a likely outcome.

Try It

Visit this page in your course online to check your understanding.

How Can Cities be Viewed As Examples of Economies of Scale?

Why are people and economic activity concentrated in cities, rather than distributed evenly across a country? The fundamental reason must be related to the idea of economies of scale—that grouping economic activity is more productive in many cases than spreading it out. For example, cities provide a large group of nearby customers, so that businesses can produce at an efficient economy of scale. They also provide a large group of workers and suppliers, so that business can hire easily and purchase whatever specialized inputs they need. Many of the attractions of cities, like sports stadiums and museums, can operate only if they can draw on a large nearby population base. Cities are big enough to offer a wide variety of products, which is what many shoppers are looking for.

These factors are not exactly economies of scale in the narrow sense of the production function of a single firm, but they are related to growth in the overall size of population and market in an area. Cities are sometimes called “agglomeration economies.” These agglomeration factors help to explain why every economy, as it develops, has an increasing proportion of its population living in urban areas. In the United States, about 80% of the population now lives in metropolitan areas (which include the suburbs around cities), compared to just 40% in 1900. However, in poorer nations of the world,
including much of Africa, the proportion of the population in urban areas is only about 30%. One of the great challenges for these countries as their economies grow will be to manage the growth of the great cities that will arise. If cities offer economic advantages that are a form of economies of scale, then why don’t all or most people live in one giant city? At some point, agglomeration economies must turn into diseconomies. For example, traffic congestion may reach a point where the gains from being geographically nearby are counterbalanced by how long it takes to travel. High densities of people, cars, and factories can mean more garbage and air and water pollution. Facilities like parks or museums may become overcrowded. There may be economies of scale for negative activities like crime, because high densities of people and businesses, combined with the greater impersonality of cities, make it easier for illegal activities as well as legal ones. The future of cities, both in the United States and in other countries around the world, will be determined by their ability to benefit from the economies of agglomeration and to minimize or counterbalance the corresponding diseconomies.

Shifting Patterns of Long-Run Average Cost

New developments in production technology can shift the long-run average cost curve in ways that can alter the size distribution of firms in an industry.

For much of the twentieth century, the most common change has been to see alterations in technology, like the assembly line or the large department store, where large-scale producers seemed to gain an advantage over smaller ones. In the long-run average cost curve, the downward-sloping economies of scale portion of the curve stretched over a larger quantity of output.

However, new production technologies do not inevitably lead to a greater average size for firms. For example, in recent years some new technologies for generating electricity on a smaller scale have appeared. The traditional coal-burning electricity plants needed to produce 300 to 600 megawatts of power to exploit economies of scale fully. However, high-efficiency turbines to produce electricity from burning natural gas can produce electricity at a competitive price while producing a smaller quantity of 100 megawatts or less. These new technologies create the possibility for smaller companies or plants to generate electricity as efficiently as large ones. Another example of a technology-driven shift to smaller plants may be taking place in the tire industry. A traditional mid-size tire plant produces about six million tires per year. However, in 2000, the Italian company Pirelli introduced a new tire factory that uses many robots. The Pirelli tire plant produced only about one million tires per year, but did so at a lower average cost than a traditional mid-sized tire plant.

Controversy has simmered in recent years over whether the new information and communications technologies will lead to a larger or smaller size for firms. On one side, the new technology may make it easier for small firms to reach out beyond their local geographic area and find customers across a state, or the nation, or even across international boundaries. This factor might seem to predict a future with a larger number of small competitors. On the other side, perhaps the new information and communications technology will create “winner-take-all” markets where one large company will tend to command a large share of total sales, as Microsoft has done in the production of software for personal computers or Amazon has done in online bookselling. Moreover, improved information and communication technologies might make it easier to manage many different plants and operations across the country or around the world, and thus encourage larger firms. This ongoing battle between the forces of smallness and largeness will be of great interest to economists, businesspeople, and policymakers.

Amazon

Traditionally, bookstores have operated in retail locations with inventories held either on the shelves or in the back of the store. These retail locations were very pricey in terms of rent. Until very recently, Amazon has had no retail locations; it sells almost entirely online and delivers by mail. Amazon offers almost any book in print, convenient purchasing, and prompt delivery by mail. Amazon holds its inventories in huge warehouses in low-rent locations around the world. The warehouses are highly computerized using robots and relatively low-skilled workers, making for low average costs per sale. Amazon demonstrates the significant advantages economies of scale can offer to a firm that exploits those economies.
PUTTING IT TOGETHER: PRODUCTION AND COSTS

The goal of this module was to explore the relationship between the inputs used in production, the resulting output and the cost of that output. Some of what you learned was detailed and complicated, but that's because production and cost issues are too. There are a near infinite number of different types of businesses and many have a unique production process. The local bakery down the street and Wonder bread don't operate the same way or at the same scale. Neither does a home baker. So each production process and costs of production are bound to be different.

What should you remember from this module?

- Any change in the production process will cause a change in production cost.
- Average and marginal costs are per unit; fixed, variable and total costs are just dollars spent on all units of output produced.
- Long-run costs are usually less than short-run costs because you have more options.
- Scale matters for some industries, but not others. The cost of producing electricity per kilowatt-hour is much less than the cost of using a home generator. But the cost of running a taxi business with 100 cabs is not much different than running the same business with 25 cabs.

Let's return to the question posed at the beginning of the module: how does an improvement in technology affect production and costs? The answer is simple. An improvement in technology usually reduces the cost of producing a given quantity of output. In other words, a firm can produce the same amount of output at lower cost or more output for the same total cost. This can be shown graphically by an upward shift in the production curves (showing more Q for the same L) and by a downward (or rightward) shift in the cost curves (showing less cost for the same Q).
MODULE 8: PERFECT COMPETITION

WHY IT MATTERS: PERFECT COMPETITION

Why analyze a firm’s decisions under conditions of perfect competition?

This module is the second in the theory of the firm and the first of three modules examining models of market structure. Market structure means, in a nutshell, how competitive or monopolistic is a particular industry. It should be clear that market structure influences how firms behave. Figure 1 shows the full spectrum of types of markets, from perfect competition with many firms, to a monopoly, with one firm controlling the marketplace.

![Figure 1. The Spectrum of Market Structure. Firms face different competitive situations. At one extreme—perfect competition—many firms are all trying to sell identical products. At the other extreme—monopoly—only one firm is selling the product, and this firm faces no competition. Monopolistic competition and oligopoly fall between the extremes of perfect competition and monopoly. Monopolistic competition is a situation with many firms selling similar, but not identical, products. Oligopoly is a situation with few firms that sell identical or similar products.]

We start by looking at the ideal model of perfect competition. This model is a bit of a head scratcher since there are, actually, very few examples of industries like this in the real world. Why then do we study it? Here’s a question for you to think about as you move through the module: what’s so perfect about perfect competition? Hint: the model has certain ideal features that you will learn and apply to the other markets.
Have you ever noticed that all the tomatoes of the same type in a farmer’s market cost about the same price? The same thing is true of roadside vegetable stands in the countryside. If one stall in a locality has tomatoes for $3 per pound, they all do. Now the price may change from week to week, but it’s always the same across the different vendors in the market. You will soon learn why this is.

There are more similarities than differences between this module and the others that follow it. What you learn in this module will carry over and help you understand the next ones, so the more effort you put into learning this one, the easier the next three modules will be.

**INTRODUCTION TO PERFECT COMPETITION**

What you’ll learn to do: describe the characteristics of perfect competition and calculate costs, including fixed, variable, average, marginal, and total costs.
Imagine the 7-year old you had a lemonade stand. It was one of several on the street. Your neighbor, Julie, also had a lemonade stand and she typically sold her lemonade for 25 cents. You figured that in order to make more money, you would charge 50 cents and steal all her customers. Sadly, everyone bought from Julie and you had no customers at all.

Welcome to the world of perfect competition. You will see in this section that because your lemonade stands were essentially identical, in order to remain in business and make any profit, you needed to be a price-taker instead of a price-maker.

PERFECT COMPETITION

Learning Objectives

- Explain the conditions and implications of a perfectly competitive market

When you were younger did you babysit, deliver papers, or mow lawns for money? If so, you faced stiff competition from other competitors who offered identical services. There was nothing to stop others from offering their services too. All of you charged the “going rate.” If you tried to charge more, your customers would simply buy from someone else. These conditions are very similar to the conditions agricultural grower face.
Growing a crop may be more difficult to start than a babysitting or lawn mowing service, but growers face the same fierce competition. In the grand scale of world agriculture, farmers face competition from thousands of others because they sell an identical product. After all, winter wheat is winter wheat. But it is relatively easy for farmers to leave the marketplace for another crop. In this case, they do not sell the family farm, they switch crops.

Take the case of the upper Midwest region of the United States—for many generations the area was called “King Wheat.” According to the United States Department of Agriculture National Agricultural Statistics Service, statistics by state, in 1997, 11.6 million acres of wheat and 780,000 acres of corn were planted in North Dakota. In the intervening 15 or so years has the mix of crops changed? Since it is relatively easy to switch crops, did farmers change what was planted as the relative crop prices changed? We will find out at module’s end.

In the meantime, let’s consider the topic of this module—the perfectly competitive market. This is a market in which entry and exit are relatively easy and competitors are “a dime a dozen.”

All businesses face two realities: no one is required to buy their products, and even customers who might want those products may buy from other businesses instead. Firms that operate in perfectly competitive markets face this reality. In this module you will learn how such firms make decisions about how much to produce, what price to charge, whether to stay in business or not, and many others. Industries differ from one another in terms of how many firms there are, how easy or difficult it is for a new firm to enter, and the type of products that are sold. This is referred to as the market structure of the industry. In this module we focus on perfect competition. However, in other modules we will examine other market structures, including monopoly, oligopoly and monopolistic competition.

What is Perfect Competition?

Firms are said to be in perfect competition when the following conditions occur: (1) the industry has many firms and many customers; (2) all firms produce identical products; (3) sellers and buyers have all relevant information to make rational decisions about the product being bought and sold; and (4) firms can enter and leave the market without any restrictions—in other words, there is free entry and exit into and out of the market.

A perfectly competitive firm is called a price taker, because the pressure of competing firms forces them to accept the prevailing equilibrium price in the market. When a wheat grower wants to know what the going price of wheat is, he or she has to go to the computer or listen to the radio to check. The market price is determined solely by supply and demand in the entire market and not the individual farmer. If a firm in a perfectly competitive market raises the price of its product by so much as a penny, it will lose all of its sales to competitors, since no rational consumer would pay a higher price for an identical product. Perfectly competitive firms, by definition, are very small players in the overall
market, so that it can increase or decrease output without noticeably affecting the overall quantity supplied and price in the market. Since they can sell all the output they want at the going market price, they never have an incentive to offer a lower price. What this means is that a perfectly competitive firm faces a horizontal demand curve at the market price, as shown in Figure 1 below.

![Figure 2. Perfectly Competitive Price. Since a perfectly competitive firm is so small relative to the market that however much output it supplies will have no effect on the market price, it can sell all it wants at the going market price. In short, a perfectly competitive firm faces a horizontal demand curve at the market price.]

A perfectly competitive market is a hypothetical extreme; however, producers in a number of industries do face many competitor firms selling highly similar goods; as a result, they must often act as price takers. Economists often use agricultural markets as an example of perfect competition. The same crops that different farmers grow are largely interchangeable. According to the United States Department of Agriculture monthly reports, in 2015, U.S. corn farmers received an average price of $6.00 per bushel. A corn farmer who attempted to sell at $7.00 per bushel, would not have found any buyers. A perfectly competitive firm will not sell below the equilibrium price either. Why should they when they can sell all they want at the higher price? Other examples of agricultural markets that operate in close to perfectly competitive markets are small roadside produce markets and small organic farmers.

Try It

Visit this page in your course online to check your understanding.

LINK IT UP

Visit this website that reveals the current value of various commodities.

This module examines how profit-seeking firms decide how much to produce in perfectly competitive markets. Such firms will analyze their costs. In the short run, the perfectly competitive firm will seek the quantity of output where profits are highest or, if profits are not possible, where losses are lowest. In this example, the “short run” refers to a situation in which firms are producing with one fixed input and incur fixed costs of production. (In the real world, firms can have many fixed inputs.)

In the long run, perfectly competitive firms will react to profits by increasing production. They will respond to losses by reducing production or exiting the market. Ultimately, a long-run equilibrium will be attained when no new firms want to enter the market and existing firms do not want to leave the market, as economic profits have been driven down to zero.

Try It

Visit this page in your course online to check your understanding.
Watch It

Watch this video for an overview on how and why firms act the way they do in a perfectly competitive market. You’ll learn about the graphs for a perfectly competitive industry and a perfectly competitive firm, then see how cost curves are used to help identify a firm’s profits. We’ll dive deeper into each of these concepts in the pages that follow.

Watch this video online: https://youtu.be/61GCogalzVc

Glossary

**market structure**: the conditions in an industry, such as number of sellers, how easy or difficult it is for a new firm to enter, and the type of products that are sold

**perfect competition**: market structure where each firm faces many competitors that sell identical products so that no firm has any market power

**price taker**: firms in a perfectly competitive market; since no firm has any market power they must take the prevailing market price as given

---

PROFIT MAXIMIZATION IN A PERFECTLY COMPETITIVE MARKET

Learning Objectives

- Determine profits and costs by comparing total revenue and total cost
- Use marginal revenue and marginal costs to find the level of output that will maximize the firm’s profits

How Perfectly Competitive Firms Make Output Decisions

A perfectly competitive firm has only one major decision to make—namely, what quantity to produce. To understand why this is so, consider the basic definition of **profit**:

\[
\text{Profit} = \text{Total revenue} - \text{Total cost} = (\text{Price}) (\text{Quantity produced}) - (\text{Average cost}) (\text{Quantity produced})
\]

Since a perfectly competitive firm must accept the price for its output as determined by the product’s market demand and supply, it cannot choose the price it charges. Rather, the perfectly competitive firm can choose to sell any quantity of output at exactly the same price. This implies that the firm faces a perfectly elastic demand curve for its product: buyers are willing to buy any number of units of output from the firm at the market price. When the perfectly competitive
firm chooses what quantity to produce, then this quantity—along with the prices prevailing in the market for output and inputs—will determine the firm’s total revenue, total costs, and ultimately, level of profits.

Determining the Highest Profit by Comparing Total Revenue and Total Cost

A perfectly competitive firm can sell as large a quantity as it wishes, as long as it accepts the prevailing market price. Total revenue is going to increase as the firm sells more, depending on the price of the product and the number of units sold. If you increase the number of units sold at a given price, then total revenue will increase. If the price of the product increases for every unit sold, then total revenue also increases.

As an example of how a perfectly competitive firm decides what quantity to produce, consider the case of a small farmer who produces raspberries and sells them frozen for $4 per pack. Sales of one pack of raspberries will bring in $4, two packs will be $8, three packs will be $12, and so on. If, for example, the price of frozen raspberries doubles to $8 per pack, then sales of one pack of raspberries will be $8, two packs will be $16, three packs will be $24, and so on.

Total revenue and total costs for the raspberry farm are shown in Table 1 and also appear in Figure 1.

<table>
<thead>
<tr>
<th>Quantity (Q)</th>
<th>Total Revenue (TR)</th>
<th>Total Cost (TC)</th>
<th>Profit</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>$0</td>
<td>$62</td>
<td>−$62</td>
</tr>
<tr>
<td>10</td>
<td>$40</td>
<td>$90</td>
<td>−$50</td>
</tr>
<tr>
<td>20</td>
<td>$80</td>
<td>$110</td>
<td>−$30</td>
</tr>
<tr>
<td>30</td>
<td>$120</td>
<td>$126</td>
<td>−$6</td>
</tr>
<tr>
<td>40</td>
<td>$160</td>
<td>$138</td>
<td>$22</td>
</tr>
<tr>
<td>50</td>
<td>$200</td>
<td>$150</td>
<td>$50</td>
</tr>
<tr>
<td>60</td>
<td>$240</td>
<td>$165</td>
<td>$75</td>
</tr>
<tr>
<td>70</td>
<td>$280</td>
<td>$190</td>
<td>$90</td>
</tr>
<tr>
<td>80</td>
<td>$320</td>
<td>$230</td>
<td>$90</td>
</tr>
<tr>
<td>90</td>
<td>$360</td>
<td>$296</td>
<td>$64</td>
</tr>
<tr>
<td>100</td>
<td>$400</td>
<td>$400</td>
<td>$0</td>
</tr>
<tr>
<td>110</td>
<td>$440</td>
<td>$550</td>
<td>−$110</td>
</tr>
<tr>
<td>120</td>
<td>$480</td>
<td>$715</td>
<td>−$235</td>
</tr>
</tbody>
</table>

In Figure 1, the horizontal axis shows the quantity of frozen raspberries produced. The vertical axis shows both total revenue and total costs, measured in dollars. The total cost curve intersects with the vertical axis at a value that shows the level of fixed costs, and then slopes upward, first at a decreasing rate, then at an increasing rate. In other words, the cost curves for a perfectly competitive firm have the same characteristics as the curves that we covered in the previous module on production and costs.

Try It
Based on its total revenue and total cost curves, a perfectly competitive firm like the raspberry farm can calculate the quantity of output that will provide the highest level of profit. At any given quantity, total revenue minus total cost will equal profit. One way to determine the most profitable quantity to produce is to see at what quantity total revenue exceeds total cost by the largest amount.

Figure 1 shows total revenue, total cost and profit using the data from Table 1. The vertical gap between total revenue and total cost is profit, for example, at Q = 60, TR = 240 and TC = 165. The difference is 75, which is the height of the profit curve at that output level. The firm doesn’t make a profit at every level of output. In this example, total costs will exceed total revenues at output levels from 0 to approximately 30, and so over this range of output, the firm will be making losses. At output levels from 40 to 100, total revenues exceed total costs, so the firm is earning profits. However, at any output greater than 100, total costs again exceed total revenues and the firm is making increasing losses. Total profits appear in the final column of Table 1. Maximum profit occurs at an output between 70 and 80, when profit equals $90.

Try It

Visit this page in your course online to check your understanding.

A higher price would mean that total revenue would be higher for every quantity sold. Graphically, the total revenue curve would be steeper, reflecting the higher price as the steeper slope. A lower price would flatten the total revenue curve, meaning that total revenue would be lower for every quantity sold. What happens if the price drops low enough so that the total revenue line is completely below the total cost curve; that is, at every level of output, total costs are higher than total revenues? In this instance, the best the firm can do is to suffer losses. However, a profit-maximizing firm will prefer the quantity of output where total revenues come closest to total costs and thus where the losses are smallest.

Comparing Marginal Revenue and Marginal Costs
The approach that we described in the previous section, using total revenue and total cost, is not the only approach to determining the profit maximizing level of output. In this section, we provide an alternative approach which uses marginal revenue and marginal cost.

Firms often do not have the necessary data they need to draw a complete total cost curve for all levels of production. They cannot be sure of what total costs would look like if they, say, doubled production or cut production in half, because they have not tried it. Instead, firms experiment. They produce a slightly greater or lower quantity and observe how it affects profits. In economic terms, this practical approach to maximizing profits means examining how changes in production affect marginal revenue and marginal cost.

As mentioned before, a firm in perfect competition faces a perfectly elastic demand curve for its product—that is, the firm’s demand curve is a horizontal line drawn at the market price level. This also means that the firm’s marginal revenue curve is the same as the firm’s demand curve. Every time a consumer demands one more unit, the firm sells one more unit and revenue increases by exactly the same amount equal to the market price. In this example, every time the firm sells a pack of frozen raspberries, the firm’s revenue increases by $4, as you can see in Table 2. This condition only holds for price taking firms in perfect competition where:

\[
\text{marginal revenue} = \text{price}
\]

The formula for marginal revenue is:

\[
\text{marginal revenue} = \frac{\text{change in total revenue}}{\text{change in quantity}}
\]

<table>
<thead>
<tr>
<th>Table 2. Marginal Revenue for Raspberries</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Price</strong></td>
</tr>
<tr>
<td>$4</td>
</tr>
<tr>
<td>$4</td>
</tr>
<tr>
<td>$4</td>
</tr>
<tr>
<td>$4</td>
</tr>
</tbody>
</table>

Notice that marginal revenue does not change as the firm produces more output. That is because the price is determined by supply and demand and does not change as the farmer produces more (keeping in mind that, due to the relative small size of each firm, increasing their supply has no impact on the total market supply where price is determined).
The equilibrium price of raspberries is determined through the interaction of market supply and market demand at $4.00.

Since a perfectly competitive firm is a price taker, it can sell whatever quantity it wishes at the market-determined price. Marginal cost, the cost per additional unit sold, is calculated by dividing the change in total cost by the change in quantity. The formula for marginal cost is:

\[
\text{marginal cost} = \frac{\text{change in total cost}}{\text{change in quantity}}
\]

Unlike marginal revenue, ordinarily, marginal cost changes as the firm produces a greater quantity of output. At first, marginal cost decreases with additional output, but then it increases with additional output. Again, note this is the same as we found in the module on production and costs.

Table 3 presents the marginal revenue and marginal costs based on the total revenue and total cost amounts introduced earlier. The marginal revenue curve shows the additional revenue gained from selling one more unit, as shown in Figure 3.

<table>
<thead>
<tr>
<th>Quantity</th>
<th>Total Revenue</th>
<th>Marginal Revenue</th>
<th>Total Cost</th>
<th>Marginal Cost</th>
<th>Profit</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>$0</td>
<td>$4</td>
<td>$62</td>
<td>–</td>
<td>-$62</td>
</tr>
<tr>
<td>10</td>
<td>$40</td>
<td>$4</td>
<td>$90</td>
<td>$2.80</td>
<td>-$50</td>
</tr>
<tr>
<td>20</td>
<td>$80</td>
<td>$4</td>
<td>$110</td>
<td>$2.00</td>
<td>-$30</td>
</tr>
<tr>
<td>30</td>
<td>$120</td>
<td>$4</td>
<td>$126</td>
<td>$1.60</td>
<td>-$6</td>
</tr>
<tr>
<td>40</td>
<td>$160</td>
<td>$4</td>
<td>$138</td>
<td>$1.20</td>
<td>$22</td>
</tr>
<tr>
<td>50</td>
<td>$200</td>
<td>$4</td>
<td>$150</td>
<td>$1.20</td>
<td>$50</td>
</tr>
<tr>
<td>60</td>
<td>$240</td>
<td>$4</td>
<td>$165</td>
<td>$1.50</td>
<td>$75</td>
</tr>
<tr>
<td>70</td>
<td>$280</td>
<td>$4</td>
<td>$190</td>
<td>$2.50</td>
<td>$90</td>
</tr>
<tr>
<td>80</td>
<td>$320</td>
<td>$4</td>
<td>$230</td>
<td>$4.00</td>
<td>$90</td>
</tr>
<tr>
<td>90</td>
<td>$360</td>
<td>$4</td>
<td>$296</td>
<td>$6.60</td>
<td>$64</td>
</tr>
<tr>
<td>Quantity</td>
<td>Total Revenue</td>
<td>Marginal Revenue</td>
<td>Total Cost</td>
<td>Marginal Cost</td>
<td>Profit</td>
</tr>
<tr>
<td>----------</td>
<td>---------------</td>
<td>------------------</td>
<td>------------</td>
<td>---------------</td>
<td>--------</td>
</tr>
<tr>
<td>100</td>
<td>$400</td>
<td>$4</td>
<td>$400</td>
<td>$10.40</td>
<td>$0</td>
</tr>
<tr>
<td>110</td>
<td>$440</td>
<td>$4</td>
<td>$550</td>
<td>$15.00</td>
<td>-$110</td>
</tr>
<tr>
<td>120</td>
<td>$480</td>
<td>$4</td>
<td>$715</td>
<td>$16.50</td>
<td>-$235</td>
</tr>
</tbody>
</table>

In the raspberry farm example, marginal cost at first declines as production increases from 10 to 20 to 30 packs of raspberries. But then marginal costs start to increase, due to diminishing marginal returns in production. If the firm is producing at a quantity where MR > MC, like 40 or 50 packs of raspberries, then it can increase profit by increasing output. The reason is since the marginal revenue exceeds the marginal cost, additional output is adding more to profit than it is taking away. If the firm is producing at a quantity where MC > MR, like 90 or 100 packs, then it can increase profit by reducing output. The firm’s profit-maximizing level of output will occur where MR = MC (or at a level close to that point).

In this example, the marginal revenue and marginal cost curves cross at a price of $4 and a quantity of 80 produced. If the farmer started out producing at a level of 60, and then experimented with increasing production to 70, marginal revenues from the increase in production would exceed marginal costs—and so profits would rise. The farmer has an incentive to keep producing. At a level of output of 80, marginal cost and marginal revenue are equal so profit doesn’t change. If the farmer then experimented further with increasing production from 80 to 90, he would find that marginal costs from the increase in production are greater than marginal revenues, and so profits would decline.

The profit-maximizing choice for a perfectly competitive firm will occur at the level of output where marginal revenue is equal to marginal cost—that is, where MR = MC. This occurs at Q = 80 in the figure.

Does Profit Maximization Occur at a Range of Output or a Specific Level of Output?

Table 1 showed that maximum profit occurs at any output level between 70 and 80 units of output. But MR = MC occurs only at 80 units of output. How can we explain this slight discrepancy? As long as MR > MC, a profit-seeking firm should keep expanding production. Expanding production into the zone where MR < MC reduces
economical profits. It's true that profit is the same at $Q = 70$ and $Q = 80$, but it's only when the firm goes beyond that that see that profits fall. Thus, $MR = MC$ is the signal to stop expanding, so that is the level of output they should target.

Because the marginal revenue received by a perfectly competitive firm is equal to the price $P$, we can also write the profit-maximizing rule for a perfectly competitive firm as a recommendation to produce at the quantity of output where $P = MC$.

**Try It**

Visit this page in your course online to check your understanding.

**Watch It**

Watch this video to practice finding the profit-maximizing point in a perfectly competitive firm. Mr. Clifford reminds us that in a perfectly competitive market, the demand curve is a horizontal line, which also happens to be the marginal revenue. You can use the acronym MR. DARP to remember that marginal revenue=demand=average revenue=price. The ideal production point is the place where MR=MC. Watch this video online: [https://youtu.be/Z9e_7j9WzA0](https://youtu.be/Z9e_7j9WzA0)

**Glossary**

**marginal revenue:** the additional revenue gained from selling one more unit of output

**profit:** the difference between total revenues and total costs

**profit-maximizing rule for a perfectly competitive firm:** produce the level of output where marginal revenue equals marginal cost

---

**Licensing & Attributions**

CC licensed content, Shared previously

- How Perfectly Competitive Firms Make Output Decisions. Authored by: OpenStax College. Located at: [https://cnx.org/contents/XAl2LLVA@7.32:EkZLadKh@7/How-Perfectly-Competitive-Firm#ch08mod02_tab01](https://cnx.org/contents/XAl2LLVA@7.32:EkZLadKh@7/How-Perfectly-Competitive-Firm#ch08mod02_tab01). License: CC BY: Attribution. License Terms: Download for free at [http://cnx.org/contents/bc498e1f-efe9-43a0-8dea-d3569ad09a82@4.4](http://cnx.org/contents/bc498e1f-efe9-43a0-8dea-d3569ad09a82@4.4)

All rights reserved content

- Perfect Competition in the Short Run- Microeconomics 3.8. Provided by: ACDC Leadership. Located at: [https://www.youtube.com/watch?v=Z9e_7j9WzA0](https://www.youtube.com/watch?v=Z9e_7j9WzA0). License: Other: License Terms: Standard YouTube License

---

**LEARN BY DOING: PROFIT MAXIMIZATION IN A PERFECTLY COMPETITIVE FIRM**
INTRODUCTION TO PROFIT IN A PERFECTLY COMPETITIVE FIRM

What you’ll learn to do: analyze a firm’s profit margin
So far, you’ve learned about perfect competition and what quantity a perfectly competitive firm will want to produce. In this section, we’ll examine profit and determine how much profit a perfectly competitive firm can earn, and at what point it should consider shutting down. Learn how perfectly competitive firms make their one important decision of how much to produce.

**CALCULATING PROFITS AND LOSSES**

**Learning Objectives**

- Describe a firm’s profit margin
- Use the average cost curve to calculate and analyze a firm’s profits and losses
- Identify and explain the firm’s break-even point

**Profits and Losses with the Average Cost Curve**

Does maximizing profit (producing where \( MR = MC \)) imply an actual economic profit? The answer depends on firm’s profit margin (or average profit), which is the relationship between price and average total cost. If the price that a firm charges is higher than its average cost of production for that quantity produced, then the firm’s profit margin is positive and it is earning economic profits. Conversely, if the price that a firm charges is lower than its average cost of production, the firm’s profit margin is negative and it is suffering an economic loss. You might think that, in this situation,
the farmer may want to shut down immediately. Remember, however, that the firm has already paid for fixed costs, such as equipment, so it may make sense to continue to produce and incur a loss. Figure 1 illustrates three situations: (a) where at the profit maximizing quantity of output (where \( P = MC \)), price is greater than average cost, (b) where at the profit maximizing quantity of output (where \( P = MC \)), price equals average cost, and (c) where at the profit maximizing quantity of output (where \( P = MC \)), price is less than average cost.

First consider a situation where the price is equal to $5 for a pack of frozen raspberries. The rule for a profit-maximizing perfectly competitive firm is to produce the level of output where Price= MR = MC, so the raspberry farmer will produce a quantity of approximately 85, which is labeled as \( E' \) in Figure 1(a). The firm’s average cost of production is labeled \( C' \). Thus, the firm’s profit margin is the distance between \( E' \) and \( C' \), and it is positive. The firm is making money, but how much? Remember that the area of a rectangle is equal to its base multiplied by its height. The farm’s total revenue at this price will be shown by the rectangle from the origin over to a quantity of 85 packs (the base) up to point \( E' \) (the height), over to the price of $5, and back to the origin. The average cost of producing 85 packs is shown by point \( C' \) or about $3.50. Total costs will be the quantity of 85 times the average cost of $3.50, which is shown by the area of the rectangle from the origin to a quantity of 90, up to point \( C \), over to the vertical axis and down to the origin. The difference between total revenues and total costs is profits. Thus, profits will be the blue shaded rectangle on top.

We calculate this as:

\[
\text{profit} = \text{total revenue} - \text{total cost} = (85) (\$5.00) - (85) (\$3.50)
\]
Or, we can calculate it as:

\[
\text{profit} = (\text{price} - \text{average cost}) \times \text{quantity}
\]

\[
= (5.00 - 3.50) \times 85
\]

\[
= 170
\]

Now consider Figure 1(b), where the price has fallen to $2.75 for a pack of frozen raspberries. Again, the perfectly competitive firm will choose the level of output where Price = MR = MC, but in this case, the quantity produced will be 75. At this price and output level, where the marginal cost curve is crossing the average cost curve, the price the firm receives is exactly equal to its average cost of production. We call this the break-even point, since the profit margin is zero.

The farm's total revenue at this price will be shown by the large shaded rectangle from the origin over to a quantity of 75 packs (the base) up to point E (the height), over to the price of $2.75, and back to the origin. The height of the average cost curve at Q = 75, i.e. point E, shows the average cost of producing this quantity. Total costs will be the quantity of 75 times the average cost of $2.75, which is shown by the area of the rectangle from the origin to a quantity of 75, up to point E, over to the vertical axis and down to the origin. It should be clear that the rectangles for total revenue and total cost are the same. Thus, the firm is making zero profit. The calculations are as follows:

\[
\text{profit} = \text{total revenue} - \text{total cost}
\]

\[
= (75)(2.75) - (75)(2.75)
\]

\[
= 0
\]

Or, we can calculate it as:

\[
\text{profit} = (\text{price} - \text{average cost}) \times \text{quantity}
\]

\[
= (2.75 - 2.75) \times 75
\]

\[
= 0
\]

In Figure 1(c), the market price has fallen still further to $2.00 for a pack of frozen raspberries. At this price, marginal revenue intersects marginal cost at a quantity of 65. The farm's total revenue at this price will be shown by the large shaded rectangle from the origin over to a quantity of 65 packs (the base) up to point E" (the height), over to the price of $2, and back to the origin. The average cost of producing 65 packs is shown by Point C" which shows the average cost of producing 50 packs is about $2.73. Since the price is less than average cost, the firm's profit margin is negative. Total costs will be the quantity of 65 times the average cost of $2.73, which the area of the rectangle from the origin to a quantity of 50, up to point C", over to the vertical axis and down to the origin shows. It should be clear from examining the two rectangles that total revenue is less than total cost. Thus, the firm is losing money and the loss (or negative profit) will be the rose-shaded rectangle.

The calculations are:

\[
\text{profit} = (\text{total revenue} - \text{total cost})
\]

\[
= (65)(2.00) - (65)(2.73)
\]

\[
= -47.45
\]

Or:

\[
\text{profit} = (\text{price} - \text{average cost}) \times \text{quantity}
\]

\[
= (2.00 - 2.73) \times 65
\]

\[
= -47.45
\]
If the market price that a perfectly competitive firm receives leads it to produce at a quantity where the price is greater than average cost, the firm will earn profits. If the price the firm receives causes it to produce at a quantity where price equals average cost, which occurs at the minimum point of the AC curve, then the firm earns zero profits. Finally, if the price the firm receives leads it to produce at a quantity where the price is less than average cost, the firm will earn losses. Table 1 summarizes this.

Table 1. Profit and Average Total Cost

<table>
<thead>
<tr>
<th>If…</th>
<th>Then…</th>
</tr>
</thead>
<tbody>
<tr>
<td>Price &gt; ATC</td>
<td>Firm earns an economic profit</td>
</tr>
<tr>
<td>Price = ATC</td>
<td>Firm earns zero economic profit</td>
</tr>
<tr>
<td>Price &lt; ATC</td>
<td>Firm earns a loss</td>
</tr>
</tbody>
</table>

Which intersection should a firm choose?

At a price of $2, MR intersects MC at two points: Q = 20 and Q = 65. It never makes sense for a firm to choose a level of output on the downward sloping part of the MC curve, because the profit is lower (the loss is bigger). Thus, the correct choice of output is Q = 65.

Watch It

Watch this video for more practice solving for the profit-maximizing point and finding total revenue using a table. Watch this video online: [https://youtu.be/BQvtnjWZ0ig](https://youtu.be/BQvtnjWZ0ig)

Try It

Play the simulation below multiple times to practice applying these concepts and to see how different choices lead to different outcomes. Visit this page in your course online to use this simulation.

Glossary

break-even point: the level of output where price just equals average total cost, so profit is zero
profit margin: at any given quantity of output, the difference between price and average total cost; also known as average profit
THE SHUTDOWN POINT

Learning Objectives

- Determine when a firm should continue producing in the short run or at which point it should shutdown

The Shutdown Point

The possibility that a firm may earn losses raises a question: why can the firm not avoid losses by shutting down and not producing at all? The answer is that shutting down can reduce variable costs to zero, but in the short run, the firm has already committed to pay its fixed costs. As a result, if the firm produces a quantity of zero, it would still make losses because it would still need to pay for its fixed costs. Therefore when a firm is experiencing losses, it must face a question: should it continue producing or should it shut down?

As an example, consider the situation of the Yoga Center, which has signed a contract to rent space that costs $10,000 per month. If the firm decides to operate, its marginal costs for hiring yoga teachers is $15,000 for the month. If the firm shuts down, it must still pay the rent, but it would not need to hire labor. Table 1 shows three possible scenarios. In the first scenario, the Yoga Center does not have any clients, and therefore does not make any revenues, in which case it faces losses of $10,000 equal to the fixed costs. In the second scenario, the Yoga Center has clients that earn the center revenues of $12,000 for the month, but ultimately experiences losses of $13,000 due to having to hire yoga instructors to cover the classes. In the third scenario, the Yoga Center earns revenues of $20,000 for the month, but experiences losses of $5,000.

In all three cases, the Yoga Center loses money. In all three cases, when the rental contract expires in the long run, assuming revenues do not improve, the firm should exit this business. In the short run, though, the decision varies depending on the level of losses and whether the firm can cover its variable costs. In scenario 1, the center does not have any revenues, so hiring yoga teachers would increase variable costs and losses, so it should shut down and only incur its fixed costs. In scenario 2, the center’s losses are greater because it does not make enough revenue to cover its variable costs, so it should shut down immediately and only incur its fixed costs. If price is below the minimum average variable cost, the firm would lose less money by shutting down. In contrast, in scenario 3 the revenue that the center can earn is high enough that the losses diminish when it remains open, so the center should remain open in the short run.

### Table 1. Should the Yoga Center Shut Down Now or Later?

<table>
<thead>
<tr>
<th>Scenario 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>If the center shuts down now, revenues are zero but it will not incur any variable costs and would only need to pay fixed costs of $10,000.</td>
</tr>
</tbody>
</table>

profit = total revenue – (fixed costs + variable cost)
profit = 0 – $10,000 = –$10,000

**Scenario 2**

The center earns revenues of $10,000, and variable costs are $15,000. The center should shut down now.

profit = total revenue – (fixed costs + variable cost)

profit = $12,000 – ($10,000 + $15,000) = –$13,000

**Scenario 3**

The center earns revenues of $20,000, and variable costs are $15,000. The center should continue in business.

profit = total revenue – (fixed costs + variable cost)

profit = $20,000 – ($10,000 + $15,000) = –$5,000

Figure 1 illustrates the lesson that remaining open requires the price to exceed the firm’s average variable cost. When the firm is operating below the break-even point, where price equals average cost, it is operating at a loss so it faces two options: continue to produce and lose money or shutdown. Which option is preferable? The one that loses the least money is the best choice.

**Try It**

Visit this page in your course online to check your understanding.

Let’s return to the raspberry farm example. At a price of $2.00 per pack, as Figure 1(a) illustrates, if the farm stays in operation it will produce at a level of 65 packs of raspberries, and it will make losses of $47.45 (as explained earlier). The alternative would be to shutdown and lose all the fixed costs of $62.00. Since losing $47.45 is preferable to losing $62.00, the profit maximizing (or in this case the loss minimizing) choice is to stay in operation. The key reason is because price is above average variable cost. This means that at the current price the farm can pay all its variable costs, and have some revenue left over to pay some of the fixed costs. So the loss represents the part of the fixed costs the farm can’t pay, which is less than the entire fixed costs. However, if the price declined to $1.50 per pack, as shown in Figure 1(b), and if the firm applied its rule of producing where P = MR = MC, it would produce a quantity of 58. This price is below average variable cost for this level of output. If the farmer cannot pay workers (the variable costs), then it has to shut down. At this price and output, total revenues would be $87 (quantity of 58 times price of $1.50) and total cost would be $162.34 (Fixed cost of $62 and variable cost of $100.34), for overall losses of $75.34. If the farm shuts down, it must pay only its fixed costs of $62, so shutting down is preferable to selling at a price of $1.50 per pack.
Figure 1. The Shutdown Point for the Raspberry Farm. In panel (a), the farm produces where \( MR = MC \) at \( Q = 65 \). It is making losses of $47.50, but price is above average variable cost, so it continues to operate. In panel (b), demand has fallen so that price ($1.50) is less than average variable cost ($1.72). If the farm continues to operate where \( MR = MC \) at \( Q = 58 \), it will lose $75.34. If it shuts down, it will owe the fixed costs of $62. The farm will lose less by shutting down.

Looking at Table 2, if the price falls below about $1.72, the minimum average variable cost, the firm must shut down.

<table>
<thead>
<tr>
<th>Quantity (Q)</th>
<th>Average Variable Cost (AVC)</th>
<th>Average Cost (AC)</th>
<th>Marginal Cost (MC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>10</td>
<td>$2.80</td>
<td>$9.00</td>
<td>$2.80</td>
</tr>
<tr>
<td>20</td>
<td>$2.40</td>
<td>$5.50</td>
<td>$2.00</td>
</tr>
<tr>
<td>30</td>
<td>$2.13</td>
<td>$4.20</td>
<td>$1.60</td>
</tr>
<tr>
<td>40</td>
<td>$1.90</td>
<td>$3.45</td>
<td>$1.20</td>
</tr>
<tr>
<td>50</td>
<td>$1.76</td>
<td>$3.00</td>
<td>$1.20</td>
</tr>
<tr>
<td>60</td>
<td>$1.72</td>
<td>$2.75</td>
<td>$1.50</td>
</tr>
<tr>
<td>70</td>
<td>$1.83</td>
<td>$2.71</td>
<td>$2.50</td>
</tr>
<tr>
<td>80</td>
<td>$2.10</td>
<td>$2.88</td>
<td>$4.00</td>
</tr>
<tr>
<td>90</td>
<td>$2.60</td>
<td>$3.29</td>
<td>$6.60</td>
</tr>
<tr>
<td>100</td>
<td>$3.38</td>
<td>$4.00</td>
<td>$10.40</td>
</tr>
<tr>
<td>110</td>
<td>$4.44</td>
<td>$5.00</td>
<td>$15.00</td>
</tr>
<tr>
<td>120</td>
<td>$5.44</td>
<td>$5.96</td>
<td>$31.50</td>
</tr>
</tbody>
</table>

The intersection of the average variable cost curve and the marginal cost curve, which shows the price below which the firm would lack enough revenue to cover its variable costs, is called the **shutdown point**. If the perfectly competitive firm faces a market price above the shutdown point, then the firm is at least covering its average variable costs. At a
price above the shutdown point, the firm is also making enough revenue to cover at least a portion of fixed costs, so it should limp ahead even if it is making losses in the short run, since at least those losses will be smaller than if the firm shuts down immediately and incurs a loss equal to total fixed costs. However, if the firm is receiving a price below the price at the shutdown point, then the firm is not even covering its variable costs. In this case, staying open is making the firm’s losses larger, and it should shut down immediately. To summarize, if:

- price < minimum average variable cost, then firm shuts down
- price > minimum average variable cost, then firm stays in business

Try It
Visit this page in your course online to check your understanding.

Short-Run Outcomes for Perfectly Competitive Firms

The average cost and average variable cost curves divide the marginal cost curve into three segments, as Figure 2 shows. At the market price, which the perfectly competitive firm accepts as given, the profit-maximizing firm chooses the output level where price or marginal revenue, which are the same thing for a perfectly competitive firm, is equal to marginal cost: \( P = MR = MC \).

First consider the upper zone, where prices are above the level where marginal cost (MC) crosses average cost (AC) at the zero profit point. At any price above that level, the firm will earn profits in the short run. If the price falls exactly on the break even point where the MC and AC curves cross, then the firm earns zero profits. If a price falls into the zone

![Figure 2. Profit, Loss, Shutdown. We can divide marginal cost curve into three zones, based on where it is crossed by the average cost and average variable cost curves. We call the point where MC crosses AC the break even point. If the firm is operating where the market price is at a level higher than the break even point, then price will be greater than average cost and the firm is earning profits. If the price is exactly at the break even point, then the firm is making zero profits. If price falls in the zone between the shutdown point and the break even point, then the firm is making losses but will continue to operate in the short run, since it is covering its variable costs, and more if price is above the shutdown point price. However, if price falls below the price at the shutdown point, then the firm will shut down immediately, since it is not even covering its variable costs.](image-url)
between the break even point, where MC crosses AC, and the shutdown point, where MC crosses AVC, the firm will be making losses in the short run—but since the firm is more than covering its variable costs, the losses are smaller than if the firm shut down immediately. Finally, consider a price at or below the shutdown point where MC crosses AVC. At any price like this one, the firm will shut down immediately, because it cannot even cover its variable costs.

**Marginal Cost and the Firm’s Supply Curve**

For a perfectly competitive firm, the marginal cost curve is identical to the firm’s supply curve starting from the minimum point on the average variable cost curve. To understand why this perhaps surprising insight holds true, first think about what the supply curve means. A firm checks the market price and then looks at its supply curve to decide what quantity to produce. Now, think about what it means to say that a firm will maximize its profits by producing at the quantity where \( P = MC \). This rule means that the firm checks the market price, and then looks at its marginal cost to determine the quantity to produce—and makes sure that the price is greater than the minimum average variable cost. In other words, the marginal cost curve above the minimum point on the average variable cost curve becomes the firm’s supply curve.

As we discussed in the module on demand and supply, many of the reasons that supply curves shift relate to underlying changes in costs. For example, a lower price of key inputs or new technologies that reduce production costs cause supply to shift to the right. In contrast, bad weather or added government regulations can add to costs of certain goods in a way that causes supply to shift to the left. We can also interpret these shifts in the firm’s supply curve as shifts of the marginal cost curve. A shift in costs of production that increases marginal costs at all levels of output—and shifts MC upward and to the left—will cause a perfectly competitive firm to produce less at any given market price. Conversely, a shift in costs of production that decreases marginal costs at all levels of output will shift MC downward and to the right and as a result, a competitive firm will choose to expand its level of output at any given price. The following Work It Out feature will walk you through an example.

**Exercise: At What Price Should the Firm Continue Producing in the Short Run?**

To determine the short-run economic condition of a firm in perfect competition, follow the steps outlined below. Use the data in the table below.

<table>
<thead>
<tr>
<th>Q</th>
<th>P</th>
<th>TFC</th>
<th>TVC</th>
<th>TC</th>
<th>AVC</th>
<th>ATC</th>
<th>MC</th>
<th>TR</th>
<th>Profits</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>$28</td>
<td>$20</td>
<td>$0</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>1</td>
<td>$28</td>
<td>$20</td>
<td>$20</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>2</td>
<td>$28</td>
<td>$20</td>
<td>$25</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>3</td>
<td>$28</td>
<td>$20</td>
<td>$35</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>4</td>
<td>$28</td>
<td>$20</td>
<td>$52</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>5</td>
<td>$28</td>
<td>$20</td>
<td>$80</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

**Step 1.** Determine the cost structure for the firm. For a given total fixed costs and variable costs, calculate total cost, average variable cost, average total cost, and marginal cost. Follow the formulas you learned previously in the module on Production. These calculations are listed below.

<table>
<thead>
<tr>
<th>Q</th>
<th>P</th>
<th>TFC</th>
<th>TVC</th>
<th>TC (TFC+TVC)</th>
<th>AVC (TVC/Q)</th>
<th>ATC (TC/Q)</th>
<th>MC (TC₂−TC₁)/(Q₂−Q₁)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>$28</td>
<td>$20</td>
<td>$0</td>
<td>$20+$0=$20</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>
Step 2. Determine the market price that the firm receives for its product. Since the firm in perfect competition is a price taker, the market price is constant. With the given price, calculate total revenue as equal to price multiplied by quantity for all output levels produced. In this example, the given price is $28. You can see that in the second column of the following table.

Table 5. Constant Prices

<table>
<thead>
<tr>
<th>Quantity</th>
<th>Price</th>
<th>Total Revenue (P x Q)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>$28</td>
<td>$28×0=$0</td>
</tr>
<tr>
<td>1</td>
<td>$28</td>
<td>$28×1=$28</td>
</tr>
<tr>
<td>2</td>
<td>$28</td>
<td>$28×2=$56</td>
</tr>
<tr>
<td>3</td>
<td>$28</td>
<td>$28×3=$84</td>
</tr>
<tr>
<td>4</td>
<td>$28</td>
<td>$28×4=$112</td>
</tr>
<tr>
<td>5</td>
<td>$28</td>
<td>$28×5=$140</td>
</tr>
</tbody>
</table>

Step 3. Calculate profits as total cost subtracted from total revenue, as Table 6 shows.

Table 6. Firm Profits

<table>
<thead>
<tr>
<th>Quantity</th>
<th>Total Revenue</th>
<th>Total Cost</th>
<th>Profits (TR–TC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>$0</td>
<td>$20</td>
<td>$0–$20=$–$20</td>
</tr>
<tr>
<td>1</td>
<td>$28</td>
<td>$40</td>
<td>$28–$40=$–$12</td>
</tr>
<tr>
<td>2</td>
<td>$56</td>
<td>$45</td>
<td>$56–$45=$11</td>
</tr>
<tr>
<td>3</td>
<td>$84</td>
<td>$55</td>
<td>$84–$55=$29</td>
</tr>
<tr>
<td>4</td>
<td>$112</td>
<td>$72</td>
<td>$112–$72=$40</td>
</tr>
<tr>
<td>5</td>
<td>$140</td>
<td>$100</td>
<td>$140–$100=$40</td>
</tr>
</tbody>
</table>
**Step 4.** To find the profit-maximizing output level, look at the Marginal Cost column (at every output level produced), as Table 7 shows, and determine where it is equal to the market price. The output level where price equals the marginal cost is the output level that maximizes profits.

<table>
<thead>
<tr>
<th>Q</th>
<th>P</th>
<th>TFC</th>
<th>TVC</th>
<th>TC</th>
<th>AVC</th>
<th>ATC</th>
<th>MC</th>
<th>TR</th>
<th>Profits</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>$28</td>
<td>$20</td>
<td>$0</td>
<td>$20</td>
<td>–</td>
<td>–</td>
<td>$0</td>
<td>–</td>
<td>–$20</td>
</tr>
<tr>
<td>1</td>
<td>$28</td>
<td>$20</td>
<td>$20</td>
<td>$40</td>
<td>$20.00</td>
<td>$40.00</td>
<td>$20</td>
<td>$28</td>
<td>–$12</td>
</tr>
<tr>
<td>2</td>
<td>$28</td>
<td>$20</td>
<td>$25</td>
<td>$45</td>
<td>$12.50</td>
<td>$22.50</td>
<td>$5</td>
<td>$56</td>
<td>$11</td>
</tr>
<tr>
<td>3</td>
<td>$28</td>
<td>$20</td>
<td>$35</td>
<td>$55</td>
<td>$11.67</td>
<td>$18.33</td>
<td>$10</td>
<td>$84</td>
<td>$29</td>
</tr>
<tr>
<td>4</td>
<td>$28</td>
<td>$20</td>
<td>$52</td>
<td>$72</td>
<td>$13.00</td>
<td>$18.00</td>
<td>$17</td>
<td>$112</td>
<td>$40</td>
</tr>
<tr>
<td>5</td>
<td>$28</td>
<td>$20</td>
<td>$80</td>
<td>$100</td>
<td>$16.40</td>
<td>$20.40</td>
<td>$28</td>
<td>$140</td>
<td>$40</td>
</tr>
</tbody>
</table>

**Step 5.** Once you have determined the profit-maximizing output level (in this case, output quantity 5), you can look at the amount of profits made (in this case, $40).

**Step 6.** If the firm is making economic losses, the firm needs to determine whether it produces the output level where price equals marginal revenue and equals marginal cost or it shuts down and only incurs its fixed costs.

**Step 7.** For the output level where marginal revenue is equal to marginal cost, check if the market price is greater than the average variable cost of producing that output level.

- If P > AVC but P < ATC, then the firm continues to produce in the short-run, making economic losses.
- If P < AVC, then the firm stops producing and only incurs its fixed costs.

In this example, the price of $28 is greater than the AVC ($16.40) of producing 5 units of output, so the firm continues producing.

**Try It**

Visit this page in your course online to check your understanding.

**Key Concepts and Summary**

As a perfectly competitive firm produces a greater quantity of output, its total revenue steadily increases at a constant rate determined by the given market price. Profits will be highest (or losses will be smallest) at the quantity of output where total revenues exceed total costs by the greatest amount (or where total revenues fall short of total costs by the smallest amount). Alternatively, profits will be highest where marginal revenue, which is price for a perfectly competitive firm, is equal to marginal cost. If the market price faced by a perfectly competitive firm is above average cost at the profit-maximizing quantity of output, then the firm is making profits. If the market price is below average cost at the profit-maximizing quantity of output, then the firm is making losses. If the market price is equal to average cost at the profit-maximizing level of output, then the firm is making zero profits.

We call the point where the marginal cost curve crosses the average cost curve, at the minimum of the average cost curve, the break-even point. If the market price that a perfectly competitive firm faces is below average variable cost at the profit-maximizing quantity of output, then the firm should shut down operations immediately. If the market price that a perfectly competitive firm faces is above average variable cost, but below average cost, then the firm should continue producing in the short run, but exit in the long run. We call the point where the marginal cost curve crosses the average variable cost curve the shutdown point.

**Watch It**
This video carefully explains the graphs you need to understand for perfect competition (it moves slowly so feel free to change the speed settings or review the parts you need most). Watch to see how profit is calculated and why firms will eventually leave at the shutdown point.

Watch this video online: https://youtu.be/V81gQroEszI

Try It

These questions allow you to get as much practice as you need, as you can click the link at the top of the first question (“Try another version of these questions”) to get a new set of questions. Practice until you feel comfortable doing the questions.

Visit this page in your course online to practice before taking the quiz.

Glossary

**break-even point**: level of output where the marginal cost curve intersects the average cost curve at the minimum point of AC; if the price is at this point, the firm is earning zero economic profits

**shutdown point**: level of output where the marginal cost curve intersects the average variable cost curve at the minimum point of AVC; if the price is below this point, the firm should shut down immediately

INTRODUCTION TO THE LONG RUN AND EFFICIENCY IN PERFECTLY COMPETITIVE MARKETS

What you’ll learn to do: describe how perfectly competitive markets adjust to long run equilibrium
Perfectly competitive markets look different in the long run than they do in the short run. In the long run, all inputs are variable, and firms may enter or exit the industry. In this section, we will explore the process by which firms in perfectly competitive markets adjust to long-run long run equilibrium. In addition, we will revisit the concept of allocative efficiency and discover how perfectly competitive firms, unlike firms in other market structures, are allocatively efficient.

ENTRY AND EXIT DECISIONS IN THE LONG RUN

Learning Objectives

- Explain how entry and exit lead to zero profits in the long run

Entry and Exit Decisions in the Long Run

The line between the short run and the long run cannot be defined precisely with a stopwatch, or even with a calendar. It varies by industry and by specific business within an industry. The distinction between the short run and the long run is therefore more technical: in the short run, firms cannot change the usage of fixed inputs, while in the long run, the firm can adjust all factors of production.
In a competitive market, profits are a red cape that incites businesses to charge. If a business is making a profit in the short run, it has an incentive to expand existing factories or to build new ones. New firms may start production, as well. When new firms come into an industry in response to high profits, it is called **entry**.

Losses are the black thundercloud that causes businesses to flee. If a business is making losses in the short run, it will either keep limping along or just shut down, depending on whether its revenues are covering its variable costs. But in the long run, firms that are facing losses will downsize, reducing their capital stock, in hopes that smaller factories and less equipment will allow them to eliminate losses. Some firms will cease production altogether. When firms leave the industry in response to a sustained pattern of losses, it is called **exit**.

### Why do firms cease to exist?

Can we say anything about what causes a firm to exit an industry? Profits are the measurement that determines whether a business stays operating or not. Individuals start businesses with the purpose of making profits. They invest their money, time, effort, and many other resources to produce and sell something that they hope will give them something in return. Unfortunately, not all businesses are successful, and many new startups eventually realize that their “business adventure” must end.

In the model of perfectly competitive firms, those that consistently cannot make money will "exit," which is a nice, bloodless word for a more painful process. When a business fails, after all, workers lose their jobs, investors lose their money, and owners and managers can lose their dreams. Many businesses fail. The U.S. Small Business Administration indicates that in 2011, 534,907 new firms “entered,” and 575,691 firms failed.

Sometimes a business fails because of poor management or workers who are not very productive, or because of tough domestic or foreign competition. Businesses also fail from a variety of causes that might best be summarized as bad luck. For example, conditions of demand and supply in the market shift in an unexpected way, so that the prices that can be charged for outputs fall or the prices that need to be paid for inputs rise. With millions of businesses in the U.S. economy, even a small fraction of them failing will affect many people—and business failures can be very hard on the workers and managers directly involved. But from the standpoint of the overall economic system, business exits are sometimes a necessary evil if a market-oriented system is going to offer a flexible mechanism for satisfying customers, keeping costs low, and inventing new products.

### How Entry and Exit Lead to Zero Profits in the Long Run

No **perfectly competitive firm** acting alone can affect the market price. However, the combination of many firms entering or exiting the market will affect overall supply in the market. In turn, a shift in supply for the market as a whole will affect the market price. Entry and exit to and from the market are the driving forces behind a process that, in the long run, pushes the price down to minimum average total costs so that all firms are earning a zero profit.

To understand how short-run profits for a perfectly competitive firm will evaporate in the long run, imagine the following situation. The market is in **long-run equilibrium**, where all firms earn zero economic profits producing the output level where \( P = MR = MC \) and \( P = AC \). No firm has the incentive to enter or leave the market. Let’s say that the product’s demand increases, and with that, the market price goes up. The existing firms in the industry are now facing a higher price than before, so they will increase production to the new output level where \( P = MR = MC \).

This will temporarily make the market price rise above the average cost curve, and therefore, the existing firms in the market will now be earning economic profits. However, these economic profits attract other firms to enter the market. Entry of many new firms causes the market supply curve to shift to the right. As the supply curve shifts to the right, the market price starts decreasing, and with that, economic profits fall for new and existing firms. As long as there are still profits in the market, entry will continue to shift supply to the right. This will stop whenever the market price is driven down to the zero-profit level, where no firm is earning economic profits.

### Try It

Visit this page in your course online to check your understanding.

### Watch It: The meaning of Zero Economic Profits
In this clip, Tyler and Alex explain why the “zero profit” can be misleading because zero profits simply mean that a firm is covering all of its cost, including enough to pay their ordinary opportunity costs and all of their labor and capital costs (meaning that they are making enough money to be satisfied). In other words, “zero profits” is what other people may call “normal profits.”

Visit this page in your course online to view this presentation.

Short-run losses will fade away by reversing this process. Say that the market is in long-run equilibrium. This time, instead, demand decreases, and with that, the market price starts falling. The existing firms in the industry are now facing a lower price than before, and as it will be below the average cost curve, they will now be making economic losses. Some firms will continue producing where the new \( P = MR = MC \), as long as they are able to cover their average variable costs. Some firms will have to shut down immediately as they will not be able to cover their average variable costs, and will then only incur their fixed costs, minimizing their losses. Exit of many firms causes the market supply curve to shift to the left. As the supply curve shifts to the left, the market price starts rising, and economic losses start to be lower. This process ends whenever the market price rises to the zero-profit level, where the existing firms are no longer losing money and are at zero profits again. Thus, while a perfectly competitive firm can earn profits in the short run, in the long run the process of entry will push down prices until they reach the zero-profit level. Conversely, while a perfectly competitive firm may earn losses in the short run, firms will not continually lose money. In the long run, firms making losses are able to escape from their fixed costs, and their exit from the market will push the price back up to the zero-profit level. In the long run, this process of entry and exit will drive the price in perfectly competitive markets to the zero-profit point at the bottom of the AC curve, where marginal cost crosses average cost.

Let's take an example of this adjustment process. Suppose the National institutes of Health publishes a study indicating that consumption of corn leads to longer lives. The demand for corn products would increase causing an increase in the market price of corn. Farmers who are already growing corn would earn positive economic profits in the short run. In the long run, farmers would increase their acreage devoted to growing corn, perhaps by reducing their acreage of wheat. The increased market supply of corn would drive the market price of corn down to the average cost of producing corn. The lower corn price would reduce the profitability of growing corn. This process would continue until corn farmers were earning zero economic profits.

Try It

Visit this page in your course online to check your understanding.

The Long-Run Adjustment for a Constant Cost Industry

Perfect competition is often the result of a constant cost industry, where there is no advantage for a firm to be large. An increase in a firm’s capital stock, simply shifts the firm’s cost curves parallel to the right. The result is a long run industry supply curve which is very elastic. The following video will explain this with two graphs: one representing a typical firm and the other representing the market (or industry as a whole).

Watch It: Constant Cost Industry

Watch this video to see how a typical firm, as well as the industry which the firm is a part of, adjust to changes in demand for the product.

Glossary

constant cost industry: an industry whose technology is such that there is no advantage to size; a large firm faces the same average costs as a small firm does.

entry: the long-run process of firms entering an industry in response to industry profits

exit: the long-run process of firms reducing production and shutting down in response to industry losses
long-run equilibrium: where all firms earn zero economic profits producing the output level where \( P = MR = MC \) and \( P = AC \)

zero economic profits: a firm is covering all of its cost, including the opportunity costs of its capital; i.e. normal accounting profits

---

LEARN BY DOING: ENTRY AND EXIT DECISIONS IN THE LONG RUN

Try It

These questions allow you to get as much practice as you need, as you can click the link at the top of the first question (“Try another version of these questions”) to get a new set of questions. Practice until you feel comfortable doing the questions.

Visit this page in your course online to practice before taking the quiz.

---

EFFICIENCY IN PERFECTLY COMPETITIVE MARKETS

Learning Objectives

- Explain why perfectly competitive firms are both productively efficient and allocatively efficient
- Compare the model of perfect competition to real-world markets

Efficiency in Perfectly Competitive Markets
When profit-maximizing firms in perfectly competitive markets combine with utility-maximizing consumers, something remarkable happens: the resulting quantities of outputs of goods and services demonstrate both productive and allocative efficiency (terms that were first introduced in the module “Choice in a World of Scarcity”).

**Productive efficiency** means producing at the lowest cost possible; in other words, producing without waste. The quantity of output supplied is on (not inside) the production possibilities frontier. In the long run in a perfectly competitive market, because of the process of entry and exit, the price in the market is equal to the minimum of the long-run average cost curve. In other words, goods are being produced and sold at the lowest possible average cost.

**Allocative efficiency** means that among the points on the production possibility frontier, the point that is chosen is socially preferred—at least in a particular and specific sense. It means that businesses supply what is demanded, neither too much nor too little. In a perfectly competitive market, price will be equal to the marginal cost of production. Think about the price that is paid for a good as a measure of the social benefit received for that good; after all, willingness to pay conveys what the good is worth to a buyer. Then think about the marginal cost of producing the good as representing not just the cost for the firm, but more broadly as the social cost of producing that good. When perfectly competitive firms follow the rule that profits are maximized by producing at the quantity where price is equal to marginal cost, they are thus ensuring that the social benefits received from producing a good are in line with the social costs of production.

To explore what is meant by **allocative efficiency**, it is useful to walk through an example. Begin by assuming that the market for wholesale flowers is perfectly competitive, and so P = MC. Now, consider what it would mean if firms in that market produced a lesser quantity of flowers. At a lesser quantity, marginal costs will not yet have increased as much, so that price will exceed marginal cost; that is, P > MC. In that situation, the benefit to society as a whole of producing additional goods, as measured by the willingness of consumers to pay for marginal units of a good, would be higher than the cost of the inputs of labor and physical capital needed to produce the marginal good. In other words, the gains to society as a whole from producing additional marginal units will be greater than the costs.

Conversely, consider what it would mean if, compared to the level of output at the allocatively efficient choice when P = MC, firms produced a greater quantity of flowers. At a greater quantity, marginal costs of production will have increased so that P < MC. In that case, the marginal costs of producing additional flowers is greater than the benefit to society as measured by what people are willing to pay. For society as a whole, since the costs are outstripping the benefits, it will make sense to produce a lower quantity of such goods.

When perfectly competitive firms maximize their profits by producing the quantity where P = MC, they also assure that the benefits to consumers of what they are buying, as measured by the price they are willing to pay, is equal to the costs to society of producing the marginal units, as measured by the marginal costs the firm must pay—and thus that allocative efficiency holds.

The statements that a perfectly competitive market in the long run will feature both productive and allocative efficiency do need to be taken with a few grains of salt. Remember, economists are using the concept of “efficiency” in a particular and specific sense, not as a synonym for “desirable in every way.” For one thing, consumers’ ability to pay reflects the income distribution in a particular society. Thus, a homeless person may have no ability to pay for housing because they have insufficient income.

Perfect competition, in the long run, is a hypothetical benchmark. For market structures such as monopoly, monopolistic competition, and oligopoly, which are more frequently observed in the real world than perfect competition, firms will not always produce at the minimum of average cost, nor will they always set price equal to marginal cost. Thus, these other competitive situations will not produce productive and allocative efficiency.

Moreover, real-world markets include many issues that are assumed away in the model of perfect competition, including pollution, inventions of new technology, poverty which may make some people unable to pay for basic necessities of life, government programs like national defense or education, discrimination in labor markets, and buyers and sellers who must deal with imperfect and unclear information. These issues are explored in other modules. However, the theoretical efficiency of perfect competition does provide a useful benchmark for comparing the issues that arise from these real-world problems.
Goodbye Wheat, Hello corn

A quick glance at the table below reveals the dramatic increase in North Dakota corn production—more than double. Taking into consideration that corn typically yields two to three times as many bushels per acre as wheat, it is obvious there has been a significant increase in bushels of corn. Why the increase in corn acreage? Converging prices.

<table>
<thead>
<tr>
<th>Year</th>
<th>Corn (thousands of acres)</th>
<th>Wheat (thousands of acres)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1997</td>
<td>780</td>
<td>11,625</td>
</tr>
<tr>
<td>2006</td>
<td>1,690</td>
<td>8,800</td>
</tr>
</tbody>
</table>

(Source: USDA National Agricultural Statistics Service)

Historically, wheat prices have been higher than corn prices, offsetting wheat’s lower yield per acre. However, in recent years wheat and corn prices have been converging. In April 2013, Agweek reported the gap was just 71 cents per bushel. As the difference in price narrowed, switching to the production of higher yield per acre of corn simply made good business sense. Erik Younggren, president of the National Association of Wheat Growers said in the Agweek article, “I don’t think we’re going to see mile after mile of waving amber fields [of wheat] anymore.” (Until wheat prices rise, we will probably be seeing field after field of tasseled corn.)

Glossary

**allocative efficiency:** when the mix of goods being produced represents the mix that society most desires

**productive efficiency:** given the available inputs and technology, it's impossible to produce more of one good without decreasing the quantity of another good that’s produced

---

License & Attributions

CC licensed content, Shared previously

- Efficiency in Perfectly Competitive Markets. **Authored by:** OpenStax College. **Provided by:** Rice University. **Located at:** https://cnx.org/contents/XAl2LLVA@7.32:zplPce7j@3/Eciency-in-Perfectly-Compet#ch08mod04_tab01. **License:** CC BY Attribution. **License Terms:** Download for free at http://cnx.org/content/col11627/latest
PUTTING IT TOGETHER: PERFECT COMPETITION

Let’s return to the questions posed at the beginning of the module. Why do all tomatoes of the same type cost the same price in a farmer’s market? The answer is because a farmer’s market or a bunch of roadside tomato stands fit the characteristics of perfect competition: many firms (or sellers at the market), all selling a similar if not identical product, where it is easy for buyers and sellers to see what everyone is charging. In this situation, if one seller charged a higher price, none of the customers would do business with that seller, since they could get an identical product from another seller at the lower price. Why would a seller charge a lower price if they can sell all their inventory by the end of the day at the going price?

By contrast, why do different gas stations on the same strip of highway charge different amounts per gallon of gasoline? Why do different pizza restaurants charge different prices for the same product, say a large one topping pizza? We’ll learn the answers to these questions in future modules.

Finally, what’s so perfect about perfect competition? The answer is that perfect competition shows markets operating at their best. Perfect competition is productively efficient, because in the long run firms produce their products as cheaply as possible (i.e. at minimum average cost). What this means in a larger context is that the economy is operating on its production possibilities frontier, rather than inside the frontier.

It’s also allocatively efficient, meaning it’s producing the optimal quantity of output (i.e. where price equals marginal cost), because that quantity maximizes total economic surplus. What this means is that it’s producing at the right point on the production possibilities frontier.

There is a lot going for the model of perfect competition. As a result, it’s the model we typically use to compare real world industries against. We’ll take a look at these real-world industries next.
MODULE 9: MONOPOLY

WHY IT MATTERS: MONOPOLY

Why analyze a firm’s profit-maximizing strategies under conditions of a monopoly?

If perfect competition is at one end of the competitive spectrum, at the other end is monopoly. Mono means one. A monoplane is an aircraft with one wing. A monocle is a single eyeglass. Monopoly is a single supplier, the only firm in an industry. Monopolies have monopoly power, which is the ability to set the market price.

As you work through this module, think about the following questions:

- What prevents a monopoly from charging an infinite price?
- What is similar about the model of monopoly compared to perfect competition?
- What is different about the model of monopoly compared to perfect competition?

There are more industries which are monopolies than are perfectly competitive, but examples of pure monopoly are still hard to find in the U.S. Google is not a monopoly. Nor is Microsoft or Amazon. Still we can learn a lot about how those firms operate by understanding the model of monopoly. Your local power company is a monopoly, but it doesn’t operate exactly the way this module explains. How do we explain this anomaly? Let’s find out.

This module begins by describing how monopolies are protected from competition, including laws that prohibit competition, technological advantages, and certain configurations of demand and supply. It then discusses how a monopoly will choose its profit-maximizing quantity to produce and what price to charge. While a monopoly must be concerned about whether consumers will purchase its products or spend their money on something altogether different, the monopolist need not worry about the actions of other competing firms producing its products. As a result, a monopoly is not a price taker like a perfectly competitive firm, but instead it has the power to choose its market price.

Licensing & Attributions

CC licensed content, Original
- Authored by: Steven Greenlaw and Lumen Learning. License: CC BY Attribution
- Monopoly Interactive. Authored by: Clark Aldrich and Lumen Learning. License: CC BY Attribution

INTRODUCTION TO MONOPOLY

What you’ll learn to do: describe characteristics of a monopoly
Businesses love the idea of a free-market economy, so why then, would a business want to become a monopolist? Just think...if a business can control the entire market for a product, then they eliminate competition and are almost guaranteed a profit. In this section, you'll learn about what a monopoly is and why firms try to eliminate their competition.

MONOPOLIES

**Learning Objectives**

- Describe the characteristics of a monopoly

**Introduction to Monopoly**

If perfect competition is a market where firms have no market power and they simply respond to the market price, monopoly is a market with no competition at all, and firms have complete market power. In the case of monopoly, one firm supplies all of the output in a market. Since a monopoly faces no significant competition, it can charge any price it wishes. While a monopoly, by definition, refers to a single firm, in practice the term is often used to describe a market in which one firm merely has a very high market share. For example, in 2013, Microsoft’s Windows operating system ran on more than 90% of the most commonly sold personal computers. This tends to be the definition that the U.S. Department of Justice uses.

Even though there are very few true monopolies in existence, we do deal with some of those few every day, often without realizing it: the U.S. Postal Service, your electric and garbage collection companies are a few examples. Some new drugs are produced by only one pharmaceutical firm—and no close substitutes for that drug may exist.

From the mid-1990s until 2004, the U.S. Department of Justice prosecuted the Microsoft Corporation for including Internet Explorer as the default web browser with its operating system. The Justice Department’s argument was that, since Microsoft possessed an
extremely high market share in the industry for operating systems, the inclusion of a free web browser constituted unfair competition to other browsers, such as Netscape Navigator (now only a browser of the past). Since nearly everyone was using Windows, including Internet Explorer eliminated the incentive for consumers to explore other browsers and made it impossible for competitors to gain a foothold in the market.

**Monopolies and History**

In spring 1773, the East India Company, a firm that, in its time, was designated “too big to fail,” was experiencing financial difficulties. To help shore up the failing firm, the British Parliament authorized the Tea Act. The act continued the tax on teas and made the East India Company the sole legal supplier of tea to the American colonies. By November, the citizens of Boston had had enough. They refused to permit the unloading of tea, citing their main complaint: “No taxation without representation.” Several newspapers, including *The Massachusetts Gazette*, warned arriving tea-bearing ships, “We are prepared, and shall not fail to pay them an unwelcome visit by The Mohawks.”

Step forward in time to 1860—the eve of the American Civil War—to another near monopoly supplier of historical significance: the U.S. cotton industry. At that time, the Southern states provided the majority of the cotton Britain imported. The South, wanting to secede from the Union, hoped to leverage Britain’s high dependency on its cotton into formal diplomatic recognition of the Confederate States of America.

This leads us to this module’s topic: a firm that controls all (or nearly all) of the supply of a good or service—a monopoly. How do monopoly firms behave in the marketplace? Do they have “power?” Does this power potentially have unintended consequences? We’ll return to this case at the end of the module to see how the tea and cotton monopolies influenced U.S. history.

**How Monopolies Form: Barriers to Entry**

Because of the lack of competition, monopolies tend to earn significant economic profits. These profits should attract vigorous competition, like the firms we discussed already in perfect competition, and yet, because of one particular characteristic of monopoly, they do not. **Barriers to entry** are the legal, technological, or market forces that discourage or prevent potential competitors from entering a market. Barriers to entry can range from the simple and easily surmountable, such as the cost of renting retail space, to the extremely restrictive. For example, there are a finite number of radio frequencies available for broadcasting. Once the rights to all of them have been purchased, no new competitors can enter the market.

In some cases, barriers to entry may lead to monopoly. In other cases, they may limit competition to a few firms. Barriers may block entry even if the firm or firms currently in the market are earning profits. Thus, in markets with significant barriers to entry, it is not true that abnormally high profits will attract new firms, and that this entry of new firms will eventually cause the price to decline so that surviving firms earn only a normal level of profit in the long run.

There are two main types of monopolies that differ in they ways they exploit barriers of entry: natural monopolies and legal monopolies. We’ll learn more about these next.

**Try It**

Visit this page in your course online to check your understanding.
Learning Objectives

- Describe and give examples of legal monopolies
- Explain how economies of scale and the control of natural resources lead to natural monopolies
- Describe and differentiate between barriers to entry

Barriers to Entry

There are two types of monopoly, based on the kinds of barriers to entry they exploit. One is legal monopoly, where laws prohibit (or severely limit) competition. The other is natural monopoly, where the barriers to entry are something other than legal prohibition.

Legal Monopoly

For some products, the government erects barriers to entry by prohibiting or limiting competition. Under U.S. law, no organization but the U.S. Postal Service is legally allowed to deliver first-class mail. Many states or cities have laws or regulations that allow households a choice of only one electric company, one water company, and one company to pick up the garbage. Most legal monopolies are considered utilities—products necessary for everyday life—that are socially beneficial to have. As a consequence, the government allows producers to become regulated monopolies, to insure that an appropriate amount of these products is provided to consumers.

Promoting Innovation

Innovation takes time and resources to achieve. Suppose a company invests in research and development and finds the cure for the common cold. In this world of near ubiquitous information, other companies could take the formula, produce the drug, and because they did not incur the costs of research and development (R&D), undercut the price of the company that discovered the drug. Given this possibility, many firms would choose not to invest in research and development, and as a result, the world would have less innovation. To prevent this from happening, the Constitution of the United States specifies in Article I, Section 8: “The Congress shall have Power . . . To Promote the Progress of Science and Useful Arts, by securing for limited Times to Authors and Inventors the Exclusive Right to their Writings and Discoveries.” Congress used this power to create the U.S. Patent and Trademark Office, as well as the U.S. Copyright Office. A patent gives the inventor the exclusive legal right to make, use, or sell the invention for a limited time. In the
United States, exclusive patent rights last for 20 years. The idea is to provide limited monopoly power so that innovative firms can recoup their investment in R&D, but then to allow other firms to produce the product more cheaply once the patent expires.

A **trademark** is an identifying symbol or name for a particular good, like Chiquita bananas, Chevrolet cars, or the Nike “swoosh” that appears on shoes and athletic gear. Roughly 1.9 million trademarks are registered with the U.S. government. A firm can renew a trademark over and over again, as long as it remains in active use.

A **copyright**, according to the U.S. Copyright Office, “is a form of protection provided by the laws of the United States for ‘original works of authorship’ including literary, dramatic, musical, architectural, cartographic, choreographic, pantomimic, pictorial, graphic, sculptural, and audiovisual creations.” No one can reproduce, display, or perform a copyrighted work without permission of the author. Copyright protection ordinarily lasts for the life of the author plus 70 years.

Roughly speaking, patent law covers inventions and copyright protects books, songs, and art. But in certain areas, like the invention of new software, it has been unclear whether patent or copyright protection should apply. There is also a body of law known as **trade secrets**. Even if a company does not have a patent on an invention, competing firms are not allowed to steal their secrets. One famous trade secret is the formula for Coca-Cola, which is not protected under copyright or patent law, but is simply kept secret by the company.

Taken together, this combination of patents, trademarks, copyrights, and trade secret law is called **intellectual property**, because it implies ownership over an idea, concept, or image, not a physical piece of property like a house or a car. Countries around the world have enacted laws to protect intellectual property, although the time periods and exact provisions of such laws vary across countries. There are ongoing negotiations, both through the World Intellectual Property Organization (WIPO) and through international treaties, to bring greater harmony to the intellectual property laws of different countries to determine the extent to which patents and copyrights in one country will be respected in other countries.

**Try It**

Visit this page in your course online to view this presentation.

---

**Natural Monopoly**

Natural monopoly occurs where the economics of an industry naturally lead to a single firm dominating the industry. Economies of scale and sole ownership (or control) of a natural resource are two common examples of natural monopoly.

A decreasing cost industry exhibits **economies of scale**, where the technology is such that the scale of operation matters, so that the long run average cost of production is lower for a large firm than for a small one. Economies of scale can combine with the size of the market to limit competition. Figure 1 presents a long-run average cost curve for the airplane manufacturing industry. It shows economies of scale up to an output of 8,000 planes per year and a price of $P_0$, then constant returns to scale from 8,000 to 20,000 planes per year, and diseconomies of scale at a quantity of production greater than 20,000 planes per year.

Now consider the market demand curve in the diagram, which intersects the long-run average cost (LRAC) curve at an output level of 6,000 planes per year and at a price $P_1$, which is higher than $P_0$. In this situation, the market has room for only one producer. If a second firm attempts to enter the market at a smaller size, say by producing a quantity of 4,000 planes, then its average costs will be higher than the existing firm, and it will be unable to compete. If the second firm attempts to enter the market at a larger size, like 8,000 planes per year, then it could produce at a lower average cost—but it could not sell all 8,000 planes that it produced because of insufficient demand in the market.
Figure 1. Economies of Scale and Natural Monopoly. In this market, the demand curve intersects the long-run average cost (LRAC) curve at its downward-sloping part. A natural monopoly occurs when the quantity demanded is less than the minimum quantity it takes to be at the bottom of the long-run average cost curve.

Natural monopolies often arise in industries where the marginal cost of adding an additional customer is very low, once the fixed costs of the overall system are in place. Once the main water pipes are laid through a neighborhood, the marginal cost of providing water service to another home is fairly low. Once electricity lines are installed through a neighborhood, the marginal cost of providing additional electrical service to one more home is very low. It would be costly and duplicative for a second water company to enter the market and invest in a whole second set of main water pipes, or for a second electricity company to enter the market and invest in a whole new set of electrical wires. These industries offer an example where, because of economies of scale, one producer can serve the entire market more efficiently than a number of smaller producers that would need to make duplicate physical capital investments.

A natural monopoly can also arise in smaller local markets for products that are difficult to transport. For example, cement production exhibits economies of scale, and the quantity of cement demanded in a local area may not be much larger than what a single plant can produce. Moreover, the costs of transporting cement over land are high, and so a cement plant in an area without access to water transportation may be a natural monopoly.

The following video presents one interesting example of a decreasing cost industry.

Watch It: Decreasing Cost industry

Watch the selected clip from this video to learn about why Dalton, Georgia is known as the “carpet capital of the world.”

Visit this page in your course online to view this presentation.

Control of a Physical Resource

Another type of natural monopoly occurs when a company has sole ownership (or majority control) of a scarce physical resource for which there are no close substitutes. In the U.S. economy, one historical example of this pattern occurred when ALCOA—the Aluminum Company of America—controlled most of the supply of bauxite, a key mineral used in making aluminum. Back in the 1930s, when ALCOA controlled most of the bauxite, other firms were simply unable to produce enough aluminum to compete.

As another example, the majority of global diamond production is controlled by DeBeers, a multi-national company that has mining and production operations in South Africa, Botswana, Namibia, and Canada. It also has exploration activities on four continents, while directing a worldwide distribution network of rough diamonds. Though in recent years they have experienced growing competition, their impact on the rough diamond market is still considerable.
Intimidating Potential Competition

Businesses have developed a number of schemes for creating barriers to entry by deterring potential competitors from entering the market. One method is known as predatory pricing, in which a firm uses the threat of sharp price cuts to discourage competition. Predatory pricing is a violation of U.S. antitrust law, but it is difficult to prove.

Consider a large airline that provides most of the flights between two particular cities. A new, small start-up airline decides to offer service between these two cities. The large airline immediately slashes prices on this route to the bone, so that the new entrant cannot make any money. After the new entrant has gone out of business, the incumbent firm can raise prices again.

After this pattern is repeated once or twice, potential new entrants may decide that it is not wise to try to compete. Small airlines often accuse larger airlines of predatory pricing: in the early 2000s, for example, ValuJet accused Delta of predatory pricing, Frontier accused United, and Reno Air accused Northwest. In late 2009, the American Booksellers Association, which represents independently owned and often smaller bookstores, accused Amazon, Wal-Mart, and Target of predatory pricing for selling new hardcover best-sellers at low prices.

In some cases, large advertising budgets can also act as a way of discouraging the competition. If the only way to launch a successful new national cola drink is to spend more than the promotional budgets of Coca-Cola and Pepsi Cola, not too many companies will try. A firmly established brand name can be difficult to dislodge.

Regulation and Deregulation of Monopolies

Government limitations on competition used to be even more common in the United States. For most of the twentieth century, only one phone company—AT&T—was legally allowed to provide local and long distance service. From the 1930s to the 1970s, one set of federal regulations limited which destinations airlines could choose to fly to and what fares they could charge; another set of regulations limited the interest rates that banks could pay to depositors; yet another specified what trucking firms could charge customers.

What products are considered utilities depends, in part, on the available technology. Fifty years ago, local and long distance telephone service was provided over wires. It did not make much sense to have multiple companies building multiple systems of wiring across towns and across the country. AT&T lost its monopoly on long distance service when the technology for providing phone service changed from wires to microwave and satellite transmission, so that multiple firms could use the same transmission mechanism. The same thing happened to local service, especially in recent years, with the growth in cellular phone systems.

The combination of improvements in production technologies and a general sense that the markets could provide services adequately led to a wave of deregulation, starting in the late 1970s and continuing into the 1990s. This wave eliminated or reduced government restrictions on the firms that could enter, the prices that could be charged, and the quantities that could be produced in many industries, including telecommunications, airlines, trucking, banking, and electricity.

Around the world, from Europe to Latin America to Africa and Asia, many governments continue to control and limit competition in what those governments perceive to be key industries, including airlines, banks, steel companies, oil companies, and telephone companies.

Try It

Visit this page in your course online to view this presentation.

Summing Up Barriers to Entry

Table 1 lists the barriers to entry that have been discussed here. This list is not exhaustive, since firms have proved to be highly creative in inventing business practices that discourage competition. When barriers to entry exist, perfect
competition is no longer a reasonable description of how an industry works. When barriers to entry are high enough, monopoly can result.

### Table 1. Barriers to Entry

<table>
<thead>
<tr>
<th>Barrier to Entry</th>
<th>Government Role?</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Natural monopoly</td>
<td>Government often responds with regulation (or ownership)</td>
<td>Water and electric companies</td>
</tr>
<tr>
<td>Control of a physical resource</td>
<td>No</td>
<td>DeBeers for diamonds</td>
</tr>
<tr>
<td>Legal monopoly</td>
<td>Yes</td>
<td>Post office, past regulation of airlines and trucking</td>
</tr>
<tr>
<td>Patent, trademark, and copyright</td>
<td>Yes, through protection of intellectual property</td>
<td>New drugs or software</td>
</tr>
<tr>
<td>Intimidating potential competitors</td>
<td>Somewhat</td>
<td>Predatory pricing; well-known brand names</td>
</tr>
</tbody>
</table>

**Watch It**

Watch this video for an overview about monopolies, including their barriers to entry and why they are problematic for market economy. Visit this page in your course online to view this presentation.

**Try It**

Visit this page in your course online to check your understanding.

**Glossary**

- **barriers to entry**: the legal, technological, or market forces that may discourage or prevent potential competitors from entering a market
- **copyright**: a form of legal protection to prevent copying, for commercial purposes, original works of authorship, including books and music
- **deregulation**: removing government controls over setting prices and quantities in certain industries
- **economies of scale**: when a firm faces decreasing long run average costs as its level of output increases
- **intellectual property**: the body of law including patents, trademarks, copyrights, and trade secret law that protect the right of inventors to produce and sell their inventions
- **legal monopoly**: legal prohibitions against competition, such as regulated monopolies and intellectual property protection
- **monopoly**: a situation in which one firm produces all of the output in a market
- **natural monopoly**: economic conditions in the industry, for example, economies of scale or control of a critical resource, that limit effective competition
- **patent**: a government rule that gives the inventor the exclusive legal right to make, use, or sell the invention for a limited time
**predatory pricing**: when an existing firm uses sharp but temporary price cuts to discourage new competition

**trade secrets**: methods of production kept secret by the producing firm

**trademark**: an identifying symbol or name for a particular good and can only be used by the firm that registered that trademark

---

**INTRODUCTION TO PROFIT AND LOSSES IN MONOPOLIES**

What you’ll learn to do: calculate and graph a monopoly’s costs, revenues, profit and losses
We know that because a monopolist controls the market for a good or service, they get more say in how much they want to produce and what price to sell it at. In this section, you’ll see how they make those decisions.

PROFIT MAXIMIZATION FOR A MONOPOLY

Learning Objectives

- Describe how a demand curve for a monopoly differs from a demand curve for a perfectly competitive firm
- Analyze total cost and total revenue curves for a monopolist
- Describe and calculate marginal revenue and marginal cost in a monopoly
- Determine the level of output the monopolist should supply and the price it should charge in order to maximize profit

Demand Curves Perceived by a Perfectly Competitive Firm and by a Monopoly

Consider a monopoly firm, comfortably surrounded by barriers to entry so that it need not fear competition from other producers. How will this monopoly choose its profit-maximizing quantity of output, and what price will it charge? Profits for the monopolist, like any firm, will be equal to total revenues minus total costs. The pattern of costs for the monopoly can be analyzed within the same framework as the costs of a perfectly competitive firm—that is, by using total cost, fixed cost, variable cost, marginal cost, average cost, and average variable cost. However, because a monopoly faces no competition, its situation and its decision process will differ from that of a perfectly competitive firm.

A perfectly competitive firm acts as a price taker. The demand curve it perceives appears in Figure 1(a). The horizontal demand curve means that, from the viewpoint of the perfectly competitive firm, it could sell either a relatively low quantity like Q₁ or a relatively high quantity like Q₉ at the market price P.
What Defines the Market?

A monopoly is a firm that sells all or nearly all of the goods and services in a given market. But what defines the “market”?

In a famous 1947 case, the federal government accused the DuPont company of having a monopoly in the cellophane market, pointing out that DuPont produced 75% of the cellophane in the United States. DuPont countered that even though it had a 75% market share in cellophane, it had less than a 20% share of the “flexible packaging materials,” which includes all other moisture-proof papers, films, and foils. In 1956, after years of legal appeals, the U.S. Supreme Court held that the broader market definition was more appropriate, and the case against DuPont was dismissed.

Questions over how to define the market continue today. True, Microsoft in the 1990s had a dominant share of the software for computer operating systems, but in the total market for all computer software and services, including everything from games to scientific programs, the Microsoft share was only about 16% in 2000. The Greyhound bus company may have a near-monopoly on the market for intercity bus transportation, but it is only a small share of the market for intercity transportation if that market includes private cars, airplanes, and railroad service. DeBeers has a monopoly in diamonds, but it is a much smaller share of the total market for precious gemstones and an even smaller share of the total market for jewelry. A small town in the country may have only one gas station: is this gas station a “monopoly,” or does it compete with gas stations that might be five, 10, or 50 miles away?

In general, if a firm produces a product without close substitutes, then the firm can be considered a monopoly producer in a single market. But if buyers have a range of similar—even if not identical—options available from other firms, then the firm is not a monopoly. Still, arguments over whether substitutes are close or not close can be controversial.

While a monopolist can charge any price for its product, that price is nonetheless constrained by demand for the firm’s product. No monopolist, even one that is thoroughly protected by high barriers to entry, can require consumers to purchase its product. Because the monopolist is the only firm in the market, its demand curve is the same as the market demand curve, which is, unlike that for a perfectly competitive firm, downward-sloping.

Figure 1 illustrates this situation. The monopolist can either choose a point like R with a low price (P_l) and high quantity (Q_h), or a point like S with a high price (P_h) and a low quantity (Q_l), or some intermediate point. Setting the price too high will result in a low quantity sold, and will not bring in much revenue. Conversely, setting the price too low may result in a high quantity sold, but because of the low price, it will not bring in much revenue either. The challenge for the monopolist is to strike a profit-maximizing balance between the price it charges and the quantity that it sells.
What is the Difference Between Perceived Demand and Market Demand?

The demand curve as perceived by a perfectly competitive firm is not the overall market demand curve for that product. However, the firm’s demand curve as perceived by a monopoly is the same as the market demand curve. The reason for the difference is that each perfectly competitive firm perceives the demand for its products in a market that includes many other firms; in effect, the demand curve perceived by a perfectly competitive firm is a tiny slice of the entire market demand curve. In contrast, a monopoly perceives demand for its product in a market where the monopoly is the only producer.

Try It

Visit this page in your course online to check your understanding.

Total Cost and Total Revenue for a Monopolist

In order to determine profits for a monopolist, we need to first identify total revenues and total costs. An example for the hypothetical HealthPill firm is shown in Figure 2.

![Figure 2. Total Revenue and Total Cost for the HealthPill Monopoly.](image)

Total revenue for the monopoly firm called HealthPill first rises, then falls. Low levels of output bring in relatively little total revenue, because the quantity is low. High levels of output bring in relatively less revenue, because the high quantity pushes down the market price. The total cost curve is upward-sloping. Profits will be highest at the quantity of output where total revenue is most above total cost. The profit-maximizing level of output is not the same as the revenue-maximizing level of output, which should make sense, because profits take costs into account and revenues do not.

Total costs for a monopolist follow the same rules as for perfectly competitive firms. In other words, total costs increase with output at an increasing rate. Total revenue, by contrast, is different from perfect competition. Since a monopolist faces a downward sloping demand curve, the only way it can sell more output is by reducing its price. Selling more output raises revenue, but lowering price reduces it. Thus, the shape of total revenue isn’t clear. Let’s explore this using the data in Table 1, which shows points along the demand curve (quantity demanded and price), and then calculates total revenue by multiplying price times quantity. (In this example, we give the output as 1, 2, 3, 4, and so on, for the sake of simplicity. If you prefer a dash of greater realism, you can imagine that the pharmaceutical company measures these output levels and the corresponding prices per 1,000 or 10,000 pills.) As Figure 2 illustrates, total revenue for a monopolist has the shape of a hill, first rising, next flattening out, and then falling.

<table>
<thead>
<tr>
<th>Quantity</th>
<th>Total Revenue ($)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1,000</td>
</tr>
<tr>
<td>2</td>
<td>2,000</td>
</tr>
<tr>
<td>3</td>
<td>3,000</td>
</tr>
<tr>
<td>4</td>
<td>4,000</td>
</tr>
<tr>
<td>5</td>
<td>5,000</td>
</tr>
<tr>
<td>6</td>
<td>6,000</td>
</tr>
<tr>
<td>7</td>
<td>7,000</td>
</tr>
</tbody>
</table>

Table 1. Total Costs and Total Revenues of HealthPill
In this example, total revenue is highest at a quantity of 6 or 7. However, the monopolist is not seeking to maximize revenue, but instead to earn the highest possible profit. In the HealthPill example in Figure 2, the highest profit will occur at the quantity where total revenue is the farthest above total cost. This looks to be somewhere in the middle of the graph, but where exactly? It is easier to see the profit maximizing level of output by using the marginal approach, to which we turn next.

Try It
Visit this page in your course online to check your understanding.

Marginal Revenue and Marginal Cost for a Monopolist

In the real world, a monopolist often does not have enough information to analyze its entire total revenues or total costs curves; after all, the firm does not know exactly what would happen if it were to alter production dramatically. But a monopolist often has fairly reliable information about how changing output by small or moderate amounts will affect its marginal revenues and marginal costs, because it has had experience with such changes over time and because modest changes are easier to extrapolate from current experience. A monopolist can use information on marginal revenue and marginal cost to seek out the profit-maximizing combination of quantity and price.

Table 2 expands Table 1 using the figures on total costs and total revenues from the HealthPill example to calculate marginal revenue and marginal cost. Recall that marginal revenue is the additional revenue the firm receives from selling one more (or a few more) units of output. Similarly, marginal cost is the additional cost the firm incurs from producing and selling one more (or a few more) units of output. This monopoly faces a typical U-shaped average cost curve and upward-sloping marginal cost curve, as shown in Figure 3.
Notice that marginal revenue is zero at a quantity of 7, and turns negative at quantities higher than 7. It may seem counterintuitive that marginal revenue could ever be zero or negative: after all, doesn’t an increase in quantity sold always mean more revenue? For a perfect competitor, each additional unit sold brought a positive marginal revenue, because marginal revenue was equal to the given market price. However, a monopolist can sell a larger quantity and see a decline in total revenue, since in order to sell more output, the monopolist must cut the price. As the quantity sold becomes higher, at some point the drop in price is proportionally more than the increase in greater quantity of sales, causing a situation where more sales bring in less revenue. In other words, marginal revenue is negative.

Figure 3. Marginal Revenue and Marginal Cost for the HealthPill Monopoly. For a monopoly like HealthPill, marginal revenue decreases as it sells additional units of output. The marginal cost curve is upward-sloping. The profit-maximizing choice for the monopoly will be to produce at the quantity where marginal revenue is equal to marginal cost: that is, MR = MC. If the monopoly produces a lower quantity, then MR > MC at those levels of output, and the firm can make higher profits by expanding output. If the firm produces at a greater quantity, then MC > MR, and the firm can make higher profits by reducing its quantity of output.

A monopolist can determine its profit-maximizing price and quantity by analyzing the marginal revenue and marginal costs of producing an extra unit. If the marginal revenue exceeds the marginal cost, then the firm can increase profit by producing one more unit of output.

For example, at an output of 4 in Figure 3, marginal revenue is 600 and marginal cost is 250, so producing this unit will clearly add to overall profits. At an output of 5, marginal revenue is 400 and marginal cost is 400, so producing this unit still means overall profits are unchanged. However, expanding output from 5 to 6 would involve a marginal revenue of
200 and a marginal cost of 850, so that sixth unit would actually reduce profits. Thus, the monopoly can tell from the marginal revenue and marginal cost that of the choices in the table, the profit-maximizing level of output is 5.

The monopoly could seek out the profit-maximizing level of output by increasing quantity by a small amount, calculating marginal revenue and marginal cost, and then either increasing output as long as marginal revenue exceeds marginal cost or reducing output if marginal cost exceeds marginal revenue. This process works without any need to calculate total revenue and total cost. Thus, a profit-maximizing monopoly should follow the rule of producing up to the quantity where marginal revenue is equal to marginal cost—that is, MR = MC. This quantity is easy to identify graphically, where MR and MC intersect.

### Maximizing Profits

If you find it counterintuitive that producing where marginal revenue equals marginal cost will maximize profits, working through the numbers will help.

**Step 1.** Remember, we define marginal cost as the change in total cost from producing a small amount of additional output.

\[
MC = \frac{\text{change in total cost}}{\text{change in quantity produced}}
\]

**Step 2.** Note that in Table 2, as output increases from 1 to 2 units, total cost increases from $1500 to $1800. As a result, the marginal cost of the second unit will be:

\[
MC = \frac{775 - 500}{1} = \$275
\]

**Step 3.** Remember that, similarly, marginal revenue is the change in total revenue from selling a small amount of additional output.

\[
MR = \frac{\text{change in total revenue}}{\text{change in quantity sold}}
\]

**Step 4.** Note that in Table 2, as output increases from 1 to 2 units, total revenue increases from $1200 to $2200. As a result, the marginal revenue of the second unit will be:

\[
MR = \frac{2200 - 1200}{1} = \$1000
\]

### Table 3. Marginal Revenue, Marginal Cost, Marginal and Total Profit

<table>
<thead>
<tr>
<th>Quantity Q</th>
<th>Marginal Revenue</th>
<th>Marginal Cost</th>
<th>Marginal Profit</th>
<th>Total Profit</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1,200</td>
<td>500</td>
<td>700</td>
<td>700</td>
</tr>
<tr>
<td>2</td>
<td>1,000</td>
<td>275</td>
<td>725</td>
<td>1,425</td>
</tr>
<tr>
<td>3</td>
<td>800</td>
<td>225</td>
<td>575</td>
<td>2,000</td>
</tr>
<tr>
<td>4</td>
<td>600</td>
<td>250</td>
<td>350</td>
<td>2,350</td>
</tr>
<tr>
<td>5</td>
<td>400</td>
<td>400</td>
<td>0</td>
<td>2,350</td>
</tr>
<tr>
<td>6</td>
<td>200</td>
<td>850</td>
<td>-650</td>
<td>1,700</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>1,500</td>
<td>-1,500</td>
<td>200</td>
</tr>
<tr>
<td>8</td>
<td>-200</td>
<td>2,400</td>
<td>-2,600</td>
<td>-2,400</td>
</tr>
</tbody>
</table>
Table 3 repeats the marginal cost and marginal revenue data from Table 2, and adds two more columns. **Marginal profit** is the profitability of each additional unit sold. We define it as marginal revenue minus marginal cost. Finally, total profit is the sum of marginal profits. As long as marginal profit is positive, producing more output will increase total profits. When marginal profit turns negative, producing more output will decrease total profits. Total profit is maximized where marginal revenue equals marginal cost. In this example, maximum profit occurs at 5 units of output.

A perfectly competitive firm will also find its profit-maximizing level of output where MR = MC. The key difference with a perfectly competitive firm is that in the case of perfect competition, marginal revenue is equal to price (MR = P), while for a monopolist, marginal revenue is not equal to the price, because changes in quantity of output affect the price.

Choosing the Price

Once the monopolist identifies the profit maximizing quantity of output, the next step is to determine the corresponding price. This is straightforward if you remember that a firm's demand curve shows the maximum price a firm can charge to sell any quantity of output. Graphically, start from the profit maximizing quantity in Figure 3, which is 5 units of output. Draw a vertical line up to the demand curve. Then read the price off the demand curve (i.e. $800).

**Watch It**

Watch the clip from this to review how a monopolist maximizes price and to see it on a graph.

Visit this page in your course online to view this presentation.

Why is a monopolist's marginal revenue always less than the price?

The marginal revenue curve for a monopolist always lies beneath the market demand curve. To understand why, think about increasing the quantity along the demand curve by one unit, so that you take one step down the demand curve to a slightly higher quantity but a slightly lower price. A demand curve is not sequential: it is not that first we sell \(Q_1\) at a higher price, and then we sell \(Q_2\) at a lower price. Rather, a demand curve is conditional: if we charge the higher price, we would sell \(Q_1\). If, instead, we charge a lower price (on all the units that we sell), we would sell \(Q_2\).

So when we think about increasing the quantity sold by one unit, marginal revenue is affected in two ways. First, we sell one additional unit at the new market price. Second, all the previous units, which could have been sold at the higher price, now sell for less. Because of the lower price on all units sold, the marginal revenue of selling a unit is less than the price of that unit—and the marginal revenue curve is below the demand curve.

**Tip:** For a straight-line demand curve, the marginal revenue curve equals price at the lowest level of output. (Graphically, MR and demand have the same vertical axis.) As output increases, marginal revenue decreases twice as fast as demand, so that the horizontal intercept of MR is halfway to the horizontal intercept of demand. You can see this in the Figure 4.
Figure 4. The Monopolist’s Marginal Revenue Curve versus Demand Curve. Because the market demand curve is conditional, the marginal revenue curve for a monopolist lies beneath the demand curve.

Try It

Visit this page in your course online to view this presentation.

Glossary

**marginal profit**: profit of one more unit of output, computed as marginal revenue minus marginal cost

LEARN BY DOING: PROFIT MAXIMIZATION FOR A MONOPOLY

Try It

These questions allow you to get as much practice as you need, as you can click the link at the top of the first question (“Try another version of these questions”) to get a new set of questions. Practice until you feel comfortable.
Learning Objectives

- Illustrate a monopoly's profits on a graph

Computing Monopoly Profits

It is straightforward to calculate profits of given numbers for total revenue and total cost. However, the size of monopoly profits can also be illustrated graphically with Figure 1, which takes the marginal cost and marginal revenue curves from the previous exhibit and adds an average cost curve and the monopolist's perceived demand curve.

Figure 1. Illustrating Profits at the HealthPill Monopoly. This figure begins with the same marginal revenue and marginal cost curves from the HealthPill monopoly from the previous page. It then adds an average cost curve and the demand curve that the monopolist faces. The HealthPill firm first chooses the quantity where MR = MC. In this example, the quantity is 5. The monopolist then decides what price to charge by looking at the demand curve it faces. For a quantity of 5, the corresponding price on the demand curve is $800. The large box, with quantity on the horizontal axis and demand (which shows the price) on the vertical axis, shows total revenue for the firm. The lighter-shaded box, which is quantity on the horizontal axis and average cost of production on the vertical axis, shows the firm's total costs. The large total revenue box minus the smaller total cost box leaves the darkly shaded box that shows total profits. Since the price charged is above average cost, the firm is earning positive profits.
The three-step process where a monopolist selects the profit-maximizing quantity to produce, decides what price to charge, and then determines total revenue, total cost and profit. These steps include:

**Step 1: The Monopolist Determines Its Profit-Maximizing Level of Output**

Since each point on a demand curve shows price and quantity, the firm can use the points on the demand curve D to calculate total revenue, and then, based on total revenue, calculate its marginal revenue curve. The profit-maximizing quantity will occur where MR = MC—or at the last possible point before marginal costs start exceeding marginal revenue. On Figure 1, MR = MC occurs at an output of 5.

**Step 2: The Monopolist Decides What Price to Charge**

The monopolist will charge what the market is willing to pay. A dotted line drawn straight up from the profit-maximizing quantity to the demand curve shows the profit-maximizing price which, in Figure 1, is $800. This price is above the average cost curve, which shows that the firm is earning profits.

**Step 3: Calculate Total Revenue, Total Cost, and Profit**

Total revenue is the overall shaded box, where the width of the box is the quantity sold and the height is the price. In Figure 1, this is 5 x $800 = $4000. The bottom part of the shaded box, which is shaded more lightly, shows total costs; that is, quantity on the horizontal axis multiplied by average cost on the vertical axis or 5 x $330 = $1650. The larger box of total revenues minus the smaller box of total costs will equal profits, which the darkly shaded box shows. Using the numbers gives $4000 – $1650 = $2350. In a perfectly competitive market, the forces of entry would erode this profit in the long run. However, a monopolist is protected by barriers to entry. In fact, one obvious sign of a possible monopoly is when a firm earns profits year after year, while doing more or less the same thing, without ever seeing increased competition eroding those profits.

In a perfectly competitive market, the forces of entry would erode this profit in the long run. But a monopolist is protected by barriers to entry. In fact, one telltale sign of a possible monopoly is when a firm earns profits year after year, while doing more or less the same thing, without ever seeing those profits eroded by increased competition.
Let’s apply these concepts to a problem. Watch the video to walk through the steps of finding total profit for a monopolist under the following conditions:

- Demand: \( P = 100 - 2q \)
- Fixed cost: 100
- Marginal cost: 20

Watch this video online: [https://youtu.be/FiQsdBWEaMI](https://youtu.be/FiQsdBWEaMI)

Try It

Visit this page in your course online to check your understanding.

Try It

These questions allow you to get as much practice as you need, as you can click the link at the top of the first question (“Try another version of these questions”) to get a new set of questions. Practice until you feel comfortable doing the questions.

Visit this page in your course online to practice before taking the quiz.

**THE INEFFICIENCY OF MONOPOLY**

**Learning Objectives**

- Explain allocative efficiency and its implications for a monopoly

**The Inefficiency of Monopoly**

Most people criticize monopolies because they charge too high a price, but what economists object to is that monopolies do not supply enough output to be allocatively efficient. To understand why a monopoly is inefficient, it is helpful to compare it with the benchmark model of perfect competition.

**Allocative efficiency** is an economic concept regarding efficiency at the social or societal level. It refers to producing the optimal quantity of some output, the quantity where the marginal benefit to society of one more unit just equals the marginal cost. The rule of profit maximization in a world of perfect competition was for each firm to produce the quantity of output where \( P = MC \). The price (\( P \)) reflects demand, and as such is a measure of how much buyers value the good, while the marginal cost (\( MC \)) is a measure of what additional units of output cost society to produce. Following this rule assures allocative efficiency. If \( P > MC \), then the marginal benefit to society (as measured by \( P \)) is greater than the marginal cost to society of producing additional units, and a greater quantity should be produced. However, in the case
of monopoly, at the profit-maximizing level of output, price is always greater than marginal cost. You can see this in Figure 1.

![Figure 1. The Allocative Inefficiency of Monopoly.](image)

Allocative Efficiency requires production at $Q_e$ where $P = MC$. A monopoly will produce less output and sell at a higher price to maximize profit at $Q_m$ and $P_m$. Thus, monopolies don’t produce enough output to be allocatively efficient.

Thus, consumers will suffer from a monopoly because it will sell a lower quantity in the market, at a higher price, than would have been the case in a perfectly competitive market.

Watch It

Watch this video to review the key concepts about monopoly, but also to learn about how monopolies are inefficient. The allocatively efficient quantity of output, or the socially optimal quantity, is where the demand equals marginal cost, but the monopoly will not produce at this point. Instead, a monopoly produces too little output at too high a cost, resulting in deadweight loss.

Watch this video online: [https://youtu.be/ZiuBWSFlfoU](https://youtu.be/ZiuBWSFlfoU)

The problem of inefficiency for monopolies often runs even deeper than these issues, and also involves incentives for efficiency over longer periods of time. There are counterbalancing incentives here. On one side, firms may strive for new inventions and new intellectual property because they want to become monopolies and earn high profits—at least for a few years until the competition catches up. In this way, monopolies may come to exist because of competitive pressures on firms. However, once a barrier to entry is in place, a monopoly that does not need to fear competition can just produce the same old products in the same old way—while still ringing up a healthy rate of profit. John Hicks, who won the Nobel Prize for economics in 1972, wrote in 1935: “The best of all monopoly profits is a quiet life.” He did not mean the comment in a complimentary way. He meant that monopolies may bank their profits and slack off on trying to please their customers.

When AT&T provided all of the local and long-distance phone service in the United States, along with manufacturing most of the phone equipment, the payment plans and types of phones did not change much. The old joke was that you could have any color phone you wanted, as long as it was black. However, in 1982, government litigation split up AT&T into a number of local phone companies, a long-distance phone company, and a phone equipment manufacturer. An explosion of innovation followed. Services like call waiting, caller ID, three-way calling, voice mail through the phone company, mobile phones, and wireless connections to the internet all became available. Companies offered a wide range of payment plans, as well. It was no longer true that all phones were black. Instead, phones came in a wide variety of shapes and colors. The end of the telephone monopoly brought lower prices, a greater quantity of services, and also a wave of innovation aimed at attracting and pleasing customers.

Try It
Glossary

**allocative efficiency**: producing the optimal quantity of some output; the quantity where the marginal benefit to society of one more unit just equals the marginal cost

---

**PRICE DISCRIMINATION AND EFFICIENCY**

**Learning Outcomes**

- Explain price discrimination and why it is an allocatively efficient outcome

**Deadweight Loss in a Monopoly**

Think about what’s wrong with a monopoly. Lay people typically say monopolies charge too high a price, but economists argue that monopolies supply too little output to be allocatively efficient. In an earlier module on the applications of supply and demand, we introduced the concepts of consumer surplus, producer surplus and social (or economic) surplus. We defined allocative efficiency as the quantity of any product that maximizes social surplus. This definition allows us to see what’s inefficient about monopoly.
Figure 1. If a firm must charge the same price to all customers, the price and quantity that will maximize profits is $P_M$ and $Q_M$, where $MR = MC$. At that quantity, since demand is greater than MC, there remain customers willing to pay more than the marginal cost of additional additional output. Thus if the firm charged addition customers, say $P_E$, the quantity sold would increase to $Q_E$, consumers would gain consumer surplus, and the firm would gain profit. The additional social surplus would be the area of the triangle below the demand curve but above the MC curve over the quantity $Q_M$ to $Q_E$.

Take a look at Figure 1. The monopoly supplies $Q_M$ and charges $P_M$. It chooses this price and quantity to maximize profits. But consider this. Is there a consumer who is willing to pay more than the marginal cost to obtain one more unit of output? The answer is yes. You can see this in Figure 1, starting at $Q_M$; now consider one more unit of output (i.e. $Q_{M+1}$). For this unit of output, the demand curve is above the marginal cost curve. That means that some consumer is willing to pay more than it would cost to produce that additional unit of output. Suppose the firm offers to sell that additional unit for $P_E$. The consumer would get the product and earn $D-P_E$ in consumer surplus. The firm would earn $P_E - MC$ in producer surplus (or profit). And the total social surplus would increase by the additional consumer surplus plus the additional producer surplus or $D - MC$.

But we don’t need to stop here. What about one more unit of output? Since $D$ is still greater than $MC$, there’s a customer who would be willing to pay more than it would cost the firm to produce one more unit of output, so both parties could benefit from producing and selling one more. Where does this end? It ends where the social surplus is maximized and where $D = MC$. That it, it ends at $Q_E$. If you think carefully, you’ll understand that $Q_E$ & $P_E$ are the quantity and price that would occur under perfect competition. In other words, since $Q_E$ maximizes social surplus, it is the most allocatively efficient quantity.

If the monopoly produces $Q_M$ and charges $P_M$, the outcome isn’t efficient. The lost social surplus due to monopoly is called a “deadweight loss,” since it is lost to society. No one captures any of that lost value.

**Price Discrimination**

It’s likely that at this point you are experiencing some cognitive dissonance. We know that a firm will maximize profits by producing the quantity of output, $Q_M$, where $MR = MC$. But we just learned that the firm can earn even more profit by producing at $Q_E$. How can both of these statements be true? The answer is **price discrimination**. Price discrimination means charging different prices to different customers for the same product. It a firm has to charge the same price to all customers, $P_M$ and $Q_M$ will maximize profits. But if it can price discriminate, it can make even more profits.
Think about when a store runs a sale. First, they charge the normal price $P_M$ and sell the normal quantity $Q_M$. Then, they run a sale and charge $P_E$ and sell $Q_E - Q_M$. Sales are an exercise in price discrimination.

Three things are necessary for effective price discrimination. First, the firm needs to have at least some market power. If it has no market power, then it can't charge different prices for different customers. Second, the firm needs to be able to sort the customers into those willing to pay a higher price and those who are no, but who would be willing to pay a lower price. Identifying these groups of customers is difficult, since customers don't usually wear signs to that effect. Often firms create a situation where customers reveal themselves. Why do movie theaters charge lower prices for matinees (i.e. showings before evening?) The answer is because usually it's young people and retired people who go to matinees because they don't have the money to pay full price. Why don't other adults go to matinees? The answer is because they tend to work during the day (or have other things to do during weekend days). The last thing necessary for price discrimination is the inability for customers to resell the product. If a grocery store charged a higher price for female shoppers and a lower price for male shoppers, smart females would simply ask their male associates to shop for them. Since the groceries can easily be transferred to (and paid for) by the females, price discrimination wouldn't work. The same strategy would not be effective with movie tickets, since the viewers couldn't “resell” the movie experience to another person. Generally, services are harder to resell than goods, so you would expect to see more price discrimination with services.

People may not like price discrimination; they may think it’s unfair. But price discrimination also provides more consumers with the product than they otherwise would be able to afford. By reducing the deadweight loss of social surplus price discrimination is more allocatively efficient.

Watch It
Watch this video to see an examples of how price discrimination is used by firms to maximize profits and efficiency.
Watch this video online: https://youtu.be/SZCV-Gm0Tt4

Try It
Visit this page in your course online to check your understanding.

Glossary

price discrimination: charging different prices to different customers for the same product.

INTRODUCTION TO MONOPOLY AND ANTITRUST POLICY

What you’ll learn to do: analyze strategies used to control monopolies
So far, we have learned three important lessons about markets: First, that competition, by providing consumers with lower prices and a variety of innovative products, is a good thing; second, that large-scale production can dramatically lower average costs; and third, that markets in the real world are rarely perfectly competitive. As a consequence, government policymakers must determine how much to intervene to balance the potential benefits of large-scale production against the potential loss of competition that can occur when businesses grow in size, especially through mergers.

For example, in 2006, AT&T and BellSouth, two telecommunications companies, wished to merge into a single firm. In the year before the merger, AT&T was the 121st largest company in the country when ranked by sales, with $44 billion in revenues and 190,000 employees. BellSouth was the 314th largest company in the country, with $21 billion in revenues and 63,000 employees.

The two companies argued that the merger would benefit consumers, who would be able to purchase better telecommunications services at a cheaper price because the newly created firm would be able to produce more efficiently by taking advantage of economies of scale and eliminating duplicate investments. However, a number of activist groups like the Consumer Federation of America and Public Knowledge expressed fears that the merger would reduce competition and lead to higher prices for consumers for decades to come. In December 2006, the federal government allowed the merger to proceed. By 2009, the new post-merger AT&T was the eighth largest company by revenues in the United States, and by that measure the largest telecommunications company in the world. Economists have spent – and will still spend – years trying to determine whether the merger of AT&T and BellSouth, as well as other smaller mergers of telecommunications companies at about this same time, helped consumers, hurt them, or did not make much difference.

This section discusses public policy issues about competition. How can economists and governments determine when mergers of large companies like AT&T and BellSouth should be allowed and when they should be blocked? The government also plays a role in policing anticompetitive behavior other than mergers, like prohibiting certain kinds of contracts that might restrict competition. In the case of natural monopoly, however, trying to preserve competition probably will not work very well, and so government will often resort to regulation of price and/or quantity of output. In recent decades, there has been a global trend toward less government intervention in the price and output decisions of businesses.

Learning Objectives

- Explain and evaluate antitrust laws and regulations
- Calculate concentration ratios and the Herfindahl-Herschman Index (HHI)
To Merge or Not TO Merge?

If you live in the United States, there is a slightly better than 50–50 chance your home is heated and cooled using natural gas. You may even use natural gas for cooking. However, those uses are not the primary uses of natural gas in the U.S. In 2012, according to the U.S. Energy Information Administration, home heating, cooling, and cooking accounted for just 18% of natural gas usage. What accounts for the rest? The greatest uses for natural gas are the generation of electric power (39%) and in industry (30%). Together these three uses for natural gas touch many areas of our lives, so why would there be any opposition to a merger of two natural gas firms? After all, a merger could mean increased efficiencies and reduced costs to people like you and me.

In October 2011, Kinder Morgan and El Paso Corporation, two natural gas firms, announced they were merging. The announcement stated the combined firm would link “nearly every major production region with markets,” cut costs by “eliminating duplication in pipelines and other assets,” and that “the savings could be passed on to consumers.”

The objection? The $21.1 billion deal would give Kinder Morgan control of more than 80,000 miles of pipeline, making the new firm the third largest energy producer in North America. As the third largest energy producer, policymakers and the public wondered whether the cost savings really would be passed on to consumers, or would the merger give Kinder Morgan a strong oligopoly position in the natural gas marketplace?

That brings us to the central question this section poses: what should the balance be between corporate size and a larger number of competitors in a marketplace? We will also consider what role the government should play in this balancing act.

Regulations for Approving Mergers

A corporate merger occurs when two formerly separate firms combine to become a single firm. When one firm purchases another, it is called an acquisition. An acquisition may not look just like a merger, since the newly purchased firm may continue to be operated under its former company name. Mergers can also be lateral, where two firms of similar sizes combine to become one. However, both mergers and acquisitions lead to two formerly separate firms being under common ownership, and so they are commonly grouped together.

Since a merger combines two firms into one, it can reduce the extent of competition between firms. Therefore, when two U.S. firms announce a merger or acquisition where at least one of the firms is above a minimum size of sales (a threshold that moves up gradually over time, and was at $76.3 million in 2015), or certain other conditions are met, they are required under law to notify the U.S. Federal Trade Commission (FTC). The left-hand panel of Figure 2(a) shows the number of mergers submitted for review to the FTC each year from 2002 to 2015. Mergers follow the business cycle, falling after the 2001 recession, peaking in 2007 as the Great Recession struck, and then rising since 2009. The right-hand panel of Figure 2(b) shows the distribution of those mergers submitted for review in 2015 as measured by the size of the transaction. It is important to remember that this total leaves out many small mergers under $50 million, which
companies only need to report in certain limited circumstances. About a third of all reported merger and acquisition transactions in 2015 exceeded $500 million, while about 15 percent exceeded $1 billion.

Figure 2. Number and Size of Mergers. (a) The number of mergers in 1999 and 2000 were relatively high compared to the annual numbers seen from 2001–2012. While 2001 and 2007 saw a high number of mergers, these were still only about half the number of mergers in 1999 and 2000. (b) In 2012, the greatest number of mergers submitted for review was for transactions between $100 and $150 million.

The laws that give government the power to block certain mergers, and even in some cases to break up large firms into smaller ones, are called antitrust laws. Before a large merger happens, the antitrust regulators at the FTC and the U.S. Department of Justice can allow the merger, prohibit it, or allow it if certain conditions are met. One common condition is that the merger will be allowed if the firm agrees to sell off certain parts. For example, in 2006, Johnson & Johnson bought the Pfizer’s “consumer health” division, which included well-known brands like Listerine mouthwash and Sudafed cold medicine. As a condition of allowing the merger, Johnson & Johnson was required to sell off six brands to other firms, including Zantac® heartburn relief medication, Cortizone anti-itch cream, and Balmex diaper rash medication, to preserve a greater degree of competition in these markets.

The U.S. government approves most proposed mergers. In a market-oriented economy, firms have the freedom to make their own choices. Private firms generally have the freedom to:

- expand or reduce production
- set the price they choose
- open new factories or sales facilities or close them
- hire workers or to lay them off
- start selling new products or stop selling existing ones

If the owners want to acquire a firm or be acquired, or to merge with another firm, this decision is just one of many that firms are free to make. In these conditions, the managers of private firms will sometimes make mistakes. They may close down a factory which, it later turns out, would have been profitable. They may start selling a product that ends up losing money. A merger between two companies can sometimes lead to a clash of corporate personalities that makes both firms worse off. But the fundamental belief behind a market-oriented economy is that firms, not governments, are in the best position to know if their actions will lead to attracting more customers or producing more efficiently.

Indeed, government regulators agree that most mergers are beneficial to consumers. As the Federal Trade Commission has noted on its website (as of November, 2013): “Most mergers actually benefit competition and consumers by allowing firms to operate more efficiently.” At the same time, the FTC recognizes, “Some [mergers] are likely to lessen competition. That, in turn, can lead to higher prices, reduced availability of goods or services, lower quality of products, and less innovation. Indeed, some mergers create a concentrated market, while others enable a single firm to raise prices.” The challenge for the antitrust regulators at the FTC and the U.S. Department of Justice is to figure out when a merger may hinder competition. This decision involves both numerical tools and some judgments that are difficult to quantify. The following section helps explain how antitrust laws came about.

What is U.S. Antitrust Law?
In the closing decades of the 1800s, many industries in the U.S. economy were dominated by a single firm that had most of the sales for the entire country. Supporters of these large firms argued that they could take advantage of economies of scale and careful planning to provide consumers with products at low prices. However, critics pointed out that when competition was reduced, these firms were free to charge more and make permanently higher profits, and that without the goading of competition, it was not clear that they were as efficient or innovative as they could be. In many cases, these large firms were organized in the legal form of a “trust,” in which a group of formerly independent firms were consolidated together by mergers and purchases, and a group of “trustees” then ran the companies as if they were a single firm. Thus, when the U.S. government passed the Sherman Antitrust Act in 1890 to limit the power of these trusts, it was called an antitrust law. In an early demonstration of the law’s power, the U.S. Supreme Court in 1911 upheld the government’s right to break up Standard Oil, which had controlled about 90% of the country’s oil refining, into 34 independent firms, including Exxon, Mobil, Amoco, and Chevron. In 1914, the Clayton Antitrust Act outlawed mergers and acquisitions (where the outcome would be to “substantially lessen competition” in an industry), price discrimination (where different customers are charged different prices for the same product), and tied sales (where purchase of one product commits the buyer to purchase some other product). Also in 1914, the Federal Trade Commission (FTC) was created to define more specifically what competition was unfair. In 1950, the Celler-Kefauver Act extended the Clayton Act by restricting vertical and conglomerate mergers. In the twenty-first century, the FTC and the U.S. Department of Justice continue to enforce antitrust laws.

---

Watch It

Watch this video for an explanation of some of the famous anti-trust cases in U.S. history.

Watch this video online: https://youtu.be/iKTrizxhHnw

---

Try It

Visit this page in your course online to view this presentation.

---

The Four-Firm Concentration Ratio

Regulators have struggled for decades to measure the degree of monopoly power in an industry. An early tool was the concentration ratio, which measures what share of the total sales in the industry are accounted for by the largest firms, typically the top four to eight firms. For an explanation of how high market concentrations can create inefficiencies in an economy, refer to Monopoly.

Say that the market for replacing broken automobile windshields in a certain city has 18 firms with the market shares shown in Table 1, where the market share is each firm’s proportion of total sales in that market. The four-firm concentration ratio is calculated by adding the market shares of the four largest firms: in this case, 16 + 10 + 8 + 6 = 40. This concentration ratio would not be considered especially high, because the largest four firms have less than half the market.

<table>
<thead>
<tr>
<th>Smooth as Glass Repair Company</th>
<th>16% of the market</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Auto Glass Doctor Company</td>
<td>10% of the market</td>
</tr>
<tr>
<td>Your Car Shield Company</td>
<td>8% of the market</td>
</tr>
<tr>
<td>Seven firms that each have 6%</td>
<td>42% of the market</td>
</tr>
<tr>
<td>Eight firms that each have 3%</td>
<td>24% of the market</td>
</tr>
</tbody>
</table>
The concentration ratio approach can help to clarify some of the fuzziness over deciding when a merger might affect competition. For instance, if two of the smallest firms in the hypothetical market for repairing automobile windshields merged, the four-firm concentration ratio would not change—which implies that there is not much worry that the degree of competition in the market has notably diminished. However, if the top two firms merged, then the four-firm concentration ratio would become 46 (that is, 26 + 8 + 6 + 6). While this concentration ratio is modestly higher, the four-firm concentration ratio would still be less than half, so such a proposed merger might barely raise an eyebrow among antitrust regulators.

The Herfindahl-Hirshman Index

A four-firm concentration ratio is a simple tool, which may reveal only part of the story. For example, consider two industries that both have a four-firm concentration ratio of 80. However, in one industry five firms each control 20% of the market, while in the other industry, the top firm holds 77% of the market and all the other firms have 1% each. Although the four-firm concentration ratios are identical, it would be reasonable to worry more about the extent of competition in the second case—where the largest firm is nearly a monopoly—than in the first.

Another approach to measuring industry concentration that can distinguish between these two cases is called the Herfindahl-Hirschman Index (HHI). The HHI, as it is often called, is calculated by summing the squares of the market share of each firm in the industry.

Calculating HHI

**Step 1.** Calculate the HHI for a monopoly with a market share of 100%. Because there is only one firm, it has 100% market share. The HHI is $100^2 = 10,000$.

**Step 2.** For an extremely competitive industry, with dozens or hundreds of extremely small competitors, the value of the HHI might drop as low as 100 or even less. Calculate the HHI for an industry with 100 firms that each have 1% of the market. In this case, the HHI is $100(1^2) = 100$.

**Step 3.** Calculate the HHI for the industry shown in Table 1. In this case, the HHI is $16^2 + 10^2 + 8^2 + 7(6^2) + 8(3^2) = 744$.

**Step 4.** Note that the HHI gives greater weight to large firms.

**Step 5.** Consider the example given earlier, comparing one industry where five firms each have 20% of the market with an industry where one firm has 77% and the other 23 firms have 1% each. The two industries have the same four-firm concentration ratio of 80. But the HHI for the first industry is $5(20^2) = 2,000$, while the HHI for the second industry is much higher at $77^2 + 23(1^2) = 5,952$.

**Step 6.** Note that the near-monopolist in the second industry drives up the HHI measure of industrial concentration.

**Step 7.** Review Table 2 which gives some examples of the four-firm concentration ratio and the HHI in various U.S. industries in 2009. (You can find market share data from multiple industry sources. Data in the table are from: Verizon (for wireless), The Wall Street Journal (for automobiles), IDC Worldwide (for computers) and the U.S. Bureau of Transportation Statistics (for airlines).)

<table>
<thead>
<tr>
<th>U.S. Industry</th>
<th>Four-Firm Ratio</th>
<th>HHI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 2. Examples of Concentration Ratios and HHIs in the U.S. Economy, 2009</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 2. Examples of Concentration Ratios and HHIs in the U.S. Economy, 2009

<table>
<thead>
<tr>
<th>Industry</th>
<th>Four-Firm Concentration Ratio</th>
<th>HHI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wireless</td>
<td>91</td>
<td>2,311</td>
</tr>
<tr>
<td>Largest five: Verizon, AT&amp;T, Sprint Nextel, T-Mobile, MetroPCS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Automobiles</td>
<td>63</td>
<td>1,121</td>
</tr>
<tr>
<td>Largest five: GM, Toyota, Ford, Honda, Chrysler</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Computers</td>
<td>74</td>
<td>1,737</td>
</tr>
<tr>
<td>Largest five: HP, Dell, Acer, Apple, Toshiba</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Airlines</td>
<td>44</td>
<td>536</td>
</tr>
<tr>
<td>Largest five: Southwest, American, Delta, United, U.S. Airways</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In the 1980s, the FTC followed these guidelines: If a merger would result in an HHI of less than 1,000, the FTC would probably approve it. If a merger would result in an HHI of more than 1,800, the FTC would probably challenge it. If a merger would result in an HHI between 1,000 and 1,800, then the FTC would scrutinize the plan and make a case-by-case decision. However, in the last several decades, the antitrust enforcement authorities have moved away from relying as heavily on measures of concentration ratios and HHIs to determine whether a merger will be allowed, and instead carried out more case-by-case analysis on the extent of competition in different industries.

Try It

Visit this page in your course online to check your understanding.

Watch It

This video explains the two main types of mergers: vertical and horizontal. Conglomerate mergers are mergers between companies with unrelated outputs. Vertical mergers are input-output mergers, in which a company buys out another company that would be involved in the production of a good. Horizontal mergers are mergers between companies that sell the same output, or the same type of product.

The video also explains the Herfindahl-Hirschman index and gives a few examples of solving for the HHI.

Watch this video online: [https://youtu.be/wlLdX0fs6hE](https://youtu.be/wlLdX0fs6hE)

New Directions for Antitrust

Both the four-firm concentration ratio and the Herfindahl-Hirschman index share some weaknesses. First, they begin from the assumption that the “market” under discussion is well-defined, and the only question is measuring how sales are divided in that market. Second, they are based on an implicit assumption that competitive conditions across industries are similar enough that a broad measure of concentration in the market is enough to make a decision about the effects of a merger. These assumptions, however, are not always correct. In response to these two problems, the antitrust regulators have been changing their approach in the last decade or two.

Defining a market is often controversial. For example, Microsoft in the early 2000s had a dominant share of the software for computer operating systems. However, in the total market for all computer software and services, including everything from games to scientific programs, the Microsoft share was only about 16% in 2000. A narrowly defined market will tend to make concentration appear higher, while a broadly defined market will tend to make it appear smaller.

There are two especially important shifts affecting how markets are defined in recent decades: one centers on technology and the other centers on globalization. In addition, these two shifts are interconnected. With the vast
improvement in communications technologies, including the development of the Internet, a consumer can order books or pet supplies from all over the country or the world. As a result, the degree of competition many local retail businesses face has increased. The same effect may operate even more strongly in markets for business supplies, where so-called “business-to-business” websites can allow buyers and suppliers from anywhere in the world to find each other.

Globalization has changed the boundaries of markets. As recently as the 1970s, it was common for measurements of concentration ratios and HHIs to stop at national borders. Now, many industries find that their competition comes from the global market. A few decades ago, three companies, General Motors, Ford, and Chrysler, dominated the U.S. auto market. By 2007, however, these three firms were making less than half of U.S. auto sales, and facing competition from well-known car manufacturers such as Toyota, Honda, Nissan, Volkswagen, Mitsubishi, and Mazda. When HHIs are calculated with a global perspective, concentration in most major industries—including cars—is lower than in a purely domestic context.

Because attempting to define a particular market can be difficult and controversial, the Federal Trade Commission has begun to look less at market share and more at the data on actual competition between businesses. For example, in February 2007, Whole Foods Market and Wild Oats Market announced that they wished to merge. These were the two largest companies in the market that the government defined as “premium natural and organic supermarket chains.” However, one could also argue that they were two relatively small companies in the broader market for all stores that sell groceries or specialty food products.

Rather than relying on a market definition, the government antitrust regulators looked at detailed evidence on profits and prices for specific stores in different cities, both before and after other competitive stores entered or exited. Based on that evidence, the Federal Trade Commission decided to block the merger. After two years of legal battles, the merger was eventually allowed in 2009 under the conditions that Whole Foods sell off the Wild Oats brand name and a number of individual stores, to preserve competition in certain local markets.

This new approach to antitrust regulation involves detailed analysis of specific markets and companies, instead of defining a market and counting up total sales. A common starting point is for antitrust regulators to use statistical tools and real-world evidence to estimate the demand curves and supply curves faced by the firms that are proposing the merger. A second step is to specify how competition occurs in this specific industry. Some possibilities include competing to cut prices, to raise output, to build a brand name through advertising, and to build a reputation for good service or high quality. With these pieces of the puzzle in place, it is then possible to build a statistical model that estimates the likely outcome for consumers if the two firms are allowed to merge. Of course, these models do require some degree of subjective judgment, and so they can become the subject of legal disputes between the antitrust authorities and the companies that wish to merge.

Glossary

**acquisition:** when one firm purchases another

**antitrust laws:** laws that give government the power to block certain mergers, and even in some cases to break up large firms into smaller ones

**concentration ratio:** an early tool to measure the degree of monopoly power in an industry; measures what share of the total sales in the industry are accounted for by the largest firms, typically the top four to eight firms

**four-firm concentration ratio:** the percentage of the total sales in the industry that are accounted for by the largest four firms

**Herfindahl-Hirschman Index (HHI):** approach to measuring market concentration by adding the square of the market share of each firm in the industry

**market share:** the percentage of total sales in the market

**merger:** when two formerly separate firms combine to become a single firm
REGULATING ANTICOMPETITIVE BEHAVIOR

Learning Objectives

- Analyze restrictive practices, including tying sales, bundling, and predatory pricing

Regulating Anticompetitive Behavior

The U.S. antitrust laws reach beyond blocking mergers that would reduce competition to include a wide array of anticompetitive practices. For example, it is illegal for competitors to form a cartel to collude to make pricing and output decisions, as if they were a monopoly firm. The Federal Trade Commission and the U.S. Department of Justice prohibit firms from agreeing to fix prices or output, rigging bids, or sharing or dividing markets by allocating customers, suppliers, territories, or lines of commerce.

In the late 1990s, for example, the antitrust regulators prosecuted an international cartel of vitamin manufacturers, including the Swiss firm Hoffman-La Roche, the German firm BASF, and the French firm Rhone-Poulenc. These firms reached agreements on how much to produce, how much to charge, and which firm would sell to which customers. The high-priced vitamins were then bought by firms like General Mills, Kellogg, Purina-Mills, and Proctor and Gamble, which pushed up the prices more. Hoffman-La Roche pleaded guilty in May 1999 and agreed both to pay a fine of $500 million and to have at least one top executive serve four months of jail time.

Under U.S. antitrust laws, monopoly itself is not illegal. If a firm has a monopoly because of a newly patented invention, for example, the law explicitly allows a firm to earn higher-than-normal profits for a time as a reward for innovation. If a firm achieves a large share of the market by producing a better product at a lower price, such behavior is not prohibited by antitrust law.

Restrictive Practices

Antitrust law includes rules against restrictive practices—practices that do not involve outright agreements to raise price or to reduce the quantity produced, but that might have the effect of reducing competition. Antitrust cases involving restrictive practices are often controversial, because they delve into specific contracts or agreements between firms that are allowed in some cases but not in others.

For example, if a product manufacturer is selling to a group of dealers who then sell to the general public it is illegal for the manufacturer to demand a minimum resale price maintenance agreement, which would require the dealers to sell for at least a certain minimum price. A minimum price contract is illegal because it would restrict competition among
dealers. However, the manufacturer is legally allowed to “suggest” minimum prices and to stop selling to dealers who regularly undercut the suggested price. If you think this rule sounds like a fairly subtle distinction, you are right.

An exclusive dealing agreement between a manufacturer and a dealer can be legal or illegal. It is legal if the purpose of the contract is to encourage competition between dealers. For example, it is legal for the Ford Motor Company to sell its cars to only Ford dealers, for General Motors to sell to only GM dealers, and so on. However, exclusive deals may also limit competition. If one large retailer obtained the exclusive rights to be the sole distributor of televisions, computers, and audio equipment made by a number of companies, then this exclusive contract would have an anticompetitive effect on other retailers.

Tying sales happen when a customer is required to buy one product only if the customer also buys a second product. Tying sales are controversial because they force consumers to purchase a product that they may not actually want or need. Further, the additional, required products are not necessarily advantageous to the customer. Suppose that to purchase a popular DVD, the store required that you also purchase a certain portable TV model. These products are only loosely related, thus there is no reason to make the purchase of one contingent on the other. Even if a customer were interested in a portable TV, the tying to a particular model prevents the customer from having the option of selecting one from the numerous types available in the market.

A related, but not identical, concept is bundling, where a firm sells two or more products as one. Bundling typically offers an advantage for consumers by allowing them to acquire multiple products or services for a better price. For example, several cable companies allow customers to buy products like cable, internet, and a phone line through a special price available through bundling. Customers are also welcome to purchase these products separately, but the price of bundling is usually more appealing.

In some cases, we can view tying sales and bundling as anticompetitive. However, in other cases they may be legal and even common. It is common for people to purchase season tickets to a sports team or a set of concerts so to guarantee tickets to the few contests or shows that are most popular and likely to sell out. Computer software manufacturers may often bundle a number of different programs, even when the buyer wants only a few. Think about the software that is included in a new computer purchase, for example.

Predatory pricing occurs when the existing firm (or firms) reacts to a new firm by dropping prices very low, until the new firm is driven out of the market, at which point the existing firm raises prices again. This pattern of pricing is aimed at deterring new firms from entering the market. However, in practice, it can be hard to figure out when pricing is predatory. Say that American Airlines is flying between two cities, and a new airline starts flying between the same two cities, at a lower price. If American Airlines cuts its price to match the new entrant, is this predatory pricing or is it just market competition at work? A commonly proposed rule is that if a firm is selling for less than its average variable cost—that is, at a price where it should be shutting down—then there is evidence for predatory pricing. However, calculating in the real world what costs are variable and what costs are fixed is often not obvious, either.

The Microsoft antitrust case embodies many of these gray areas in restrictive practices, as the next section shows.

DID MICROSOFT® ENGAGE IN ANTICOMPETITIVE AND RESTRICTIVE PRACTICES?

The most famous restrictive practices case of recent years was a series of lawsuits by the U.S. government against Microsoft—lawsuits that were encouraged by some of Microsoft’s competitors. All sides admitted that Microsoft’s Windows program had a near-monopoly position in the market for the software used in general computer operating systems. All sides agreed that the software had many satisfied customers. All sides agreed that the capabilities of computer software that was compatible with Windows—both software produced by Microsoft and that produced by other companies—had expanded dramatically in the 1990s. Having a monopoly or a near-monopoly is not necessarily illegal in and of itself, but in cases where one company controls a great deal of the market, antitrust regulators look at any allegations of restrictive practices with special care.

The antitrust regulators argued that Microsoft had gone beyond profiting from its software innovations and its dominant position in the software market for operating systems, and had tried to use its market power in operating systems software to take over other parts of the software industry. For example, the government argued that
Microsoft had engaged in an anticompetitive form of exclusive dealing by threatening computer makers that, if they did not leave another firm’s software off their machines (specifically, Netscape’s Internet browser), then Microsoft would not sell them its operating system software. Microsoft was accused by the government antitrust regulators of tying together its Windows operating system software, where it had a monopoly, with its Internet Explorer browser software, where it did not have a monopoly, and thus using this bundling as an anticompetitive tool. Microsoft was also accused of a form of predatory pricing; namely, giving away certain additional software products for free as part of Windows, as a way of driving out the competition from other makers of software.

In April 2000, a federal court held that Microsoft’s behavior had crossed the line into unfair competition, and recommended that the company be broken into two competing firms. However, that penalty was overturned on appeal, and in November 2002 Microsoft reached a settlement with the government that it would end its restrictive practices.

The concept of restrictive practices is continually evolving, as firms seek new ways to earn profits and government regulators define what is permissible and what is not. A situation where the law is evolving and changing is always somewhat troublesome, since laws are most useful and fair when firms know what they are in advance. In addition, since the law is open to interpretation, competitors who are losing out in the market can accuse successful firms of anticompetitive restrictive practices, and try to win through government regulation what they have failed to accomplish in the market. Officials at the Federal Trade Commission and the Department of Justice are, of course, aware of these issues, but there is no easy way to resolve them.

Try It
Visit this page in your course online to view this presentation.

Glossary

bundling: a situation in which multiple products are sold as one

exclusive dealing: an agreement that a dealer will sell only products from one manufacturer

minimum resale price maintenance agreement: an agreement that requires a dealer who buys from a manufacturer to sell for at least a certain minimum price

restrictive practices: practices that reduce competition but that do not involve outright agreements between firms to raise prices or to reduce the quantity produced

tying sales: a situation where a customer is allowed to buy one product only if the customer also buys another product

REGULATING NATURAL MONOPOLIES
Regulating Natural Monopolies

Most true monopolies today in the U.S. are regulated, natural monopolies. A natural monopoly poses a difficult challenge for competition policy, because the structure of costs and demand seems to make competition unlikely or costly. A natural monopoly arises when average costs are declining over the range of production that satisfies market demand. This typically happens when fixed costs are large relative to variable costs. As a result, one firm is able to supply the total quantity demanded in the market at lower cost than two or more firms—so splitting up the natural monopoly would raise the average cost of production and force customers to pay more.

Public utilities, the companies that have traditionally provided water and electrical service across much of the United States, are leading examples of natural monopoly. It would make little sense to argue that a local water company should be broken up into several competing companies, each with its own separate set of pipes and water supplies. Installing four or five identical sets of pipes under a city, one for each water company, so that each household could choose its own water provider, would be terribly costly. The same argument applies to the idea of having many competing companies for delivering electricity to homes, each with its own set of wires. Before the advent of wireless phones, the argument also applied to the idea of many different phone companies, each with its own set of phone wires running through the neighborhood.

The Choices in Regulating a Natural Monopoly

So what then is the appropriate competition policy for a natural monopoly? Figure 1 illustrates the case of natural monopoly, with a market demand curve that cuts through the downward-sloping portion of the average cost curve. Points A, B, C, and F illustrate four of the main choices for regulation.

![Figure 1. Regulatory Choices in Dealing with Natural Monopoly.](image)

A natural monopoly will maximize profits by producing at the quantity where marginal revenue (MR) equals marginal costs (MC) and by then looking to the market demand curve to see what price to charge for this quantity. This monopoly will produce at point A, with a quantity of 4 and a price of 9.3. If antitrust regulators split this company exactly in half, then each half would produce at point B, with average costs of 9.75 and output of 2. The regulators might require the firm to produce where marginal cost crosses the market demand curve at point C. However, if the firm is required to produce at a quantity of 6 and sell at a price of 3.5, the firm will suffer from losses. The most likely choice is point F, where the firm is required to produce a quantity of 6 and charge a price of 6.5.

Table 1 outlines the regulatory choices for dealing with a natural monopoly.
Table 1. Regulatory Choices in Dealing with Natural Monopoly

<table>
<thead>
<tr>
<th>Quantity</th>
<th>Price</th>
<th>Total Revenue*</th>
<th>Marginal Revenue</th>
<th>Total Cost</th>
<th>Marginal Cost</th>
<th>Average Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>14.7</td>
<td>14.7</td>
<td>–</td>
<td>11.0</td>
<td>–</td>
<td>11.00</td>
</tr>
<tr>
<td>2</td>
<td>12.4</td>
<td>24.7</td>
<td>10.0</td>
<td>19.5</td>
<td>8.5</td>
<td>9.75</td>
</tr>
<tr>
<td>3</td>
<td>10.6</td>
<td>31.7</td>
<td>7.0</td>
<td>25.5</td>
<td>6.0</td>
<td>8.50</td>
</tr>
<tr>
<td>4</td>
<td>9.3</td>
<td>37.2</td>
<td>5.5</td>
<td>31.0</td>
<td>5.5</td>
<td>7.75</td>
</tr>
<tr>
<td>5</td>
<td>8.0</td>
<td>40.0</td>
<td>2.8</td>
<td>35.0</td>
<td>4.0</td>
<td>7.00</td>
</tr>
<tr>
<td>6</td>
<td>6.5</td>
<td>39.0</td>
<td>–1.0</td>
<td>39.0</td>
<td>4.0</td>
<td>6.50</td>
</tr>
<tr>
<td>7</td>
<td>5.0</td>
<td>35.0</td>
<td>–4.0</td>
<td>42.0</td>
<td>3.0</td>
<td>6.00</td>
</tr>
<tr>
<td>8</td>
<td>3.5</td>
<td>28.0</td>
<td>–7.0</td>
<td>45.5</td>
<td>3.5</td>
<td>5.70</td>
</tr>
<tr>
<td>9</td>
<td>2.0</td>
<td>18.0</td>
<td>–10.0</td>
<td>49.5</td>
<td>4.0</td>
<td>5.50</td>
</tr>
</tbody>
</table>

*Total Revenue is given by multiplying price and quantity. However, some of the price values in this table have been rounded for ease of presentation.

The first possibility is to leave the natural monopoly alone. In this case, the monopoly will follow its normal approach to maximizing profits. It determines the quantity where MR = MC, which happens at point P at a quantity of 4. The firm then looks to point A on the demand curve to find that it can charge a price of 9.3 for that profit-maximizing quantity. Since the price is above the average cost curve, the natural monopoly would earn economic profits.

A second outcome arises if antitrust authorities decide to divide the company, so that the new firms can compete. As a simple example, imagine that the company is cut in half. Thus, instead of one large firm producing a quantity of 4, two half-size firms each produce a quantity of 2. Because of the declining average cost curve (AC), the average cost of production for each of the half-size companies each producing 2, as shown at point B, would be 9.75, while the average cost of production for a larger firm producing 4 would only be 7.75. Thus, the economy would become less productively efficient, since the good is being produced at a higher average cost. In a situation with a downward-sloping average cost curve, two smaller firms will always have higher average costs of production than one larger firm for any quantity of total output. In addition, the antitrust authorities must worry that splitting the natural monopoly into pieces may be only the start of their problems. If one of the two firms grows larger than the other, it will have lower average costs and may be able to drive its competitor out of the market. Alternatively, two firms in a market may discover subtle ways of coordinating their behavior and keeping prices high. Either way, the result will not be the greater competition that was desired.

A third alternative is that regulators may decide to set prices and quantities produced for this industry. The regulators will try to choose a point along the market demand curve that benefits both consumers and the broader social interest. Point C illustrates one tempting choice: the regulator requires that the firm produce the quantity of output where marginal cost crosses the demand curve at an output of 8, and charge the price of 3.5, which is equal to marginal cost at that point. This rule is appealing because it requires price to be set equal to marginal cost, which is what would occur in a perfectly competitive market, and it would assure consumers a higher quantity and lower price than at the monopoly choice A. In fact, efficient allocation of resources would occur at point C, since the value to the consumers of the last unit bought and sold in this market is equal to the marginal cost of producing it.

Attempting to bring about point C through force of regulation, however, runs into a severe difficulty. At point C, with an output of 8, a price of 3.5 is below the average cost of production, which is 5.7, and so if the firm charges a price of 3.5, it will be suffering losses. Unless the regulators or the government offer the firm an ongoing public subsidy (and there are numerous political problems with that option), the firm will lose money and go out of business.

Perhaps the most plausible option for the regulator is point F; that is, to set the price where AC crosses the demand curve at an output of 6 and a price of 6.5. This plan makes some sense at an intuitive level: let the natural monopoly charge enough to cover its average costs and earn a normal rate of profit, so that it can continue operating, but prevent the firm from raising prices and earning abnormally high monopoly profits, as it would at the monopoly choice A. Of
course, determining this level of output and price with the political pressures, time constraints, and limited information of the real world is much harder than identifying the point on a graph. For more on the problems that can arise from a centrally determined price, see the discussion of price floors and price ceilings in Demand and Supply.

Try It
Visit this page in your course online to check your understanding.

Cost-Plus versus Price Cap Regulation

Indeed, regulators of public utilities for many decades followed the general approach of attempting to choose a point like F in Figure 1. They calculated the average cost of production for the water or electricity companies, added in an amount for the normal rate of profit the firm should expect to earn, and set the price for consumers accordingly. This method was known as cost-plus regulation.

Cost-plus regulation raises difficulties of its own. If producers are reimbursed for their costs, plus a bit more, then at a minimum, producers have less reason to be concerned with high costs—because they can just pass them along in higher prices. Worse, firms under cost-plus regulation even have an incentive to generate high costs by building huge factories or employing lots of staff, because what they can charge is linked to the costs they incur.

Thus, in the 1980s and 1990s, some regulators of public utilities began to use price cap regulation, where the regulator sets a price that the firm can charge over the next few years. A common pattern was to require a price that declined slightly over time. If the firm can find ways of reducing its costs more quickly than the price caps, it can make a high level of profits. However, if the firm cannot keep up with the price caps or suffers bad luck in the market, it may suffer losses. A few years down the road, the regulators will then set a new series of price caps based on the firm’s performance.

Price cap regulation requires delicacy. It will not work if the price regulators set the price cap unrealistically low. It may not work if the market changes dramatically so that the firm is doomed to incurring losses no matter what it does—say, if energy prices rise dramatically on world markets, then the company selling natural gas or heating oil to homes may not be able to meet price caps that seemed reasonable a year or two ago. But if the regulators compare the prices with producers of the same good in other areas, they can, in effect, pressure a natural monopoly in one area to compete with the prices being charged in other areas. Moreover, the possibility of earning greater profits or experiencing losses—instead of having an average rate of profit locked in every year by cost-plus regulation—can provide the natural monopoly with incentives for efficiency and innovation.

With natural monopoly, market competition is unlikely to take root, so if consumers are not to suffer the high prices and restricted output of an unrestricted monopoly, government regulation will need to play a role. In attempting to design a system of price cap regulation with flexibility and incentive, government regulators do not have an easy task.

Watch It
Watch this video to analyze the cost curves for a natural monopoly and to consider various options for regulation. Watch this video online: https://youtu.be/lZfbZDK0hLw

Try It
Visit this page in your course online to check your understanding.

Try It
These questions allow you to get as much practice as you need, as you can click the link at the top of the first question (“Try another version of these questions”) to get a new set of questions. Practice until you feel comfortable
Regulating Natural Monopolies

Regulatory Frameworks

- **Cost-plus Regulation**: When regulators permit a regulated firm to cover its costs and make a normal level of profit.
- **Natural Monopoly**: Economic conditions in the industry where economies of scale limit effective competition.
- **Price Cap Regulation**: When the regulator sets a price a firm cannot exceed over the next few years.

Learning Objectives

- Evaluate the effectiveness of price regulation and antitrust policy.

Doubts about Regulation of Prices and Quantities

Governments at all levels across the United States have regulated prices in a wide range of industries. In some cases, like water and electricity with natural monopoly characteristics, there is some room in economic theory for such regulation. But once politicians are given a basis to intervene in markets and choose prices and quantities, it is hard to know where to stop.

Beginning in the 1970s, it became clear to policymakers of all political leanings that the existing price regulation was not working well. The United States carried out a great policy experiment—removing government controls over prices and quantities produced in airlines, railroads, trucking, intercity bus travel, natural gas, and bank interest rates. The next section discusses the outcome of deregulation in one industry in particular—airlines.

WHAT ARE THE RESULTS OF AIRLINE Deregulation?

Why did the pendulum swing in favor of deregulation? Consider the airline industry. In the early days of air travel, no airline could make a profit just by flying passengers. Airlines needed something else to carry and the Postal Service provided that something with airmail. And so the first U.S. government regulation of the airline industry happened through the Postal Service, when in 1926 the Postmaster General began giving airlines permission to fly certain routes based on the needs of mail delivery—and the airlines took some passengers along for the ride. In 1934, the Postmaster General was charged by the antitrust authorities and the Civil Aeronautics Board (CAB) was created to regulate airfares and routes instead. For 40 years, from 1938 to 1978, the CAB approved all fares, controlled all entry and exit, and
specified which airlines could fly which routes. There was zero entry of new airlines on the main routes across the country for 40 years, because the CAB did not think it was necessary.

In 1978, the Airline Deregulation Act took the government out of the business of determining airfares and schedules. The new law shook up the industry. Famous old airlines like Pan American, Eastern, and Braniff went bankrupt and disappeared. Some new airlines like People Express were created—and then vanished.

The greater competition from deregulation reduced airfares by about one-third over the next two decades, saving consumers billions of dollars a year. The average flight used to take off with just half its seats full; now it is two-thirds full, which is far more efficient. Airlines have also developed hub-and-spoke systems, where planes all fly into a central hub city at a certain time and then depart. As a result, one can fly between any of the spoke cities with just one connection—and there is greater service to more cities than before deregulation. With lower fares and more service, the number of air passengers doubled from the late 1970s to the start of the 2000s—an increase that, in turn, doubled the number of jobs in the airline industry. Meanwhile, with the watchful oversight of government safety inspectors, commercial air travel has continued to get safer over time.

The U.S. airline industry is far from perfect. For example, a string of mergers in recent years has raised concerns over how competition might be compromised.

One difficulty with government price regulation is what economists call regulatory capture, in which the firms supposedly being regulated end up playing a large role in setting the regulations that they will follow. When the airline industry was being regulated, for example, it suggested appointees to the regulatory board, sent lobbyists to argue with the board, provided most of the information on which the board made decisions, and offered well-paid jobs to at least some of the people leaving the board. In this situation, consumers can easily end up being not very well represented by the regulators. The result of regulatory capture is that government price regulation can often become a way for existing competitors to work together to reduce output, keep prices high, and limit competition.

The Effects of Deregulation

Deregulation, both of airlines and of other industries, has its negatives. The greater pressure of competition led to entry and exit. When firms went bankrupt or contracted substantially in size, they laid off workers who had to find other jobs. Market competition is, after all, a full-contact sport.

A number of major accounting scandals involving prominent corporations such as Enron, Tyco International, and WorldCom led to the Sarbanes-Oxley Act in 2002. Sarbanes-Oxley was designed to increase confidence in financial information provided by public corporations to protect investors from accounting fraud.

The Great Recession which began in late 2007 and which the U.S. economy is still struggling to recover from was caused at least in part by a global financial crisis, which began in the United States. The key component of the crisis was the creation and subsequent failure of several types of unregulated financial assets, such as collateralized mortgage obligations (CMOs, a type of mortgage-backed security), and credit default swaps (CDSs, insurance contracts on assets like CMOs that provided a payoff even if the holder of the CDS did not own the CMO). Many of these assets were rated very safe by private credit rating agencies such as Standard & Poors, Moody’s, and Fitch.

The collapse of the markets for these assets precipitated the financial crisis and led to the failure of Lehman Brothers, a major investment bank, numerous large commercial banks, such as Wachovia, and even the Federal National Mortgage Corporation (Fannie Mae), which had to be nationalized—that is, taken over by the federal government. One response to the financial crisis was the Dodd-Frank Act, which attempted major reforms of the financial system. The legislation’s purpose, as noted on dodd-frank.com is:

To promote the financial stability of the United States by improving accountability and transparency in the financial system, to end “too big to fail,” to protect the American taxpayer by ending bailouts, [and] to protect consumers from abusive financial services practices. . .

You will explore the financial crisis and the Great Recession in more detail when you study macroeconomics, but for now it should be clear that many Americans have grown disenchanted with deregulation, at least of financial markets.

All market-based economies operate against a background of laws and regulations, including laws about enforcing contracts, collecting taxes, and protecting health and the environment. The government policies discussed in this module—like blocking certain anticompetitive mergers, ending restrictive practices, imposing price cap regulation on natural monopolies, and deregulation—demonstrate the role of government to strengthen the incentives that come with a greater degree of competition.
To MERge or Not TO Merge?

So what did the Federal Trade Commission (FTC) decide on the Kinder Morgan/El Paso Corporation merger that we mentioned earlier? After careful examination, federal officials decided there was only one area of significant overlap that might provide the merged firm with strong market power. The FTC approved the merger, provided Kinder Morgan divest itself of the overlap area. Tallgrass purchased Kinder Morgan Interstate Gas Transmission, Trailblazer Pipeline Co. LLC, two processing facilities in Wyoming, and Kinder Morgan’s 50 percent interest in the Rockies Express Pipeline to meet the FTC requirements. The FTC was attempting to strike a balance between potential cost reductions resulting from economies of scale and concentration of market power.

Did the price of natural gas decrease? Yes, rather significantly. In 2010, the wellhead price of natural gas was $4.48 per thousand cubic foot; in 2012 the price had fallen to just $2.66. Was the merger responsible for the large drop in price? The answer is uncertain. The larger contributor to the sharp drop in price was the overall increase in the supply of natural gas. More and more natural gas was able to be recovered by fracturing shale deposits, a process called fracking. Fracking, which is controversial for environmental reasons, enabled the recovery of known reserves of natural gas that previously were not economically feasible to tap. Kinder Morgan’s control of 80,000-plus miles of pipeline likely made moving the gas from wellheads to end users smoother and allowed for an even greater benefit from the increased supply.

Try It

Visit this page in your course online to check your understanding.

Glossary

**Dodd-Frank Act**: legislation designed to protect consumers and end bailouts to lead to greater economic stability

**regulatory capture**: when the supposedly regulated firms end up playing a large role in setting the regulations that they will follow and as a result, they “capture” the people usually through the promise of a job in that “regulated” industry once their term in government has ended

**Sarbanes-Oxley Act**: legislation designed to increase confidence in financial information provided by public corporations to protect investors from accounting fraud

PUTTING IT TOGETHER: MONOPOLY

The goal of this module was analyze a firm’s profit maximizing strategies under conditions of monopoly. Among other things, you learned how to:

- **Describe the characteristics of a monopoly.** A monopoly occurs when a single firm supplies the whole market for some product. Because they face no direct competition, monopolies can charge any price they want and earn
economic profits, even in the long run.

- **Explain the sources of barriers to entry.** Barriers to entry are economic or legal prohibitions on other firms entering an industry to capture some of the monopoly's profits.

- **Calculate and graph a monopoly's fixed, variable, average, marginal and total costs.** Costs are computed and cost curves graphed the same way as in perfect competition. This is one of the similarities across market structures.

- **Explain why a monopoly is inefficient.** Allocative inefficiency occurs when firms produce less than the optimal supply, which monopolies do to allow them to charge a higher price. Deadweight loss is the loss in total surplus (producer + consumer surplus) that occurs at output less than the optimal one.

- **Analyze different strategies to control monopolies.** Governments use laws and regulation to reduce the inefficiency of monopolies. Regulated monopolies supply more than they otherwise would in return for a guaranteed profit.

So why are monopolies bad? It's not primarily because they charge too high a price. Rather, it's because they are allocatively inefficient, in other words, they produce too little of the product, and because (often) they are productively inefficient, that is, they don't produce as cheaply as possible. Many of the issues we explored in this module will show up again in the next module on Monopolistic Competition and Oligopoly.

Nearly all monopolies in the U.S. are regulated monopolies, meaning the prices they charge have to be reviewed and approved (or not) by a regulatory branch of the government. A good example of this is the U.S. Postal Service, whose rates must be accepted by the Postal Rate Commission. Thus, regulated monopolies don't behave exactly the way pure monopolies do, as explained in this module. Why do we study them then? Because nearly all firms in the real world have some market power, that is the ability to influence the market price. Monopoly power, the ability to set the market price, is the ultimate in market power. Understanding how monopolies exploit their power helps us understand how real world, but not-quite-monopoly firms, operate.

---

### Monopolies and History

At the beginning of the module, we presented the East India Company and the Confederate States as a monopoly or near monopoly provider of a good. Nearly every American schoolchild knows the result of the “unwelcome visit” the “Mohawks” bestowed upon Boston Harbor’s tea-bearing ships—the Boston Tea Party. Regarding the cotton industry, we also know Great Britain remained neutral during the Civil War, taking neither side during the conflict.

Did the monopoly nature of these business have unintended and historical consequences? Might the American Revolution have been deterred, if the East India Company had sailed the tea-bearing ships back to England? Might the southern states have made different decisions had they not been so confident “King Cotton” would force diplomatic recognition of the Confederate States of America? Of course, it is not possible to definitively answer these questions. We cannot roll back the clock and try a different scenario. We can, however, consider the monopoly nature of these businesses and the roles they played and hypothesize about what might have occurred under different circumstances.

Perhaps if there had been legal free tea trade, the colonists would have seen things differently. There was smuggled Dutch tea in the colonial market. If the colonists had been able to freely purchase Dutch tea, they would have paid lower prices and avoided the tax.

What about the cotton monopoly? With one in five jobs in Great Britain depending on Southern cotton and the Confederate States as nearly the sole provider of that cotton, why did Great Britain remain neutral during the Civil War? At the beginning of the war, Britain simply drew down massive stores of cotton. These stockpiles lasted until near the end of 1862. Why did Britain not recognize the Confederacy at that point? Two reasons: The Emancipation Proclamation and new sources of cotton. Having outlawed slavery throughout the United Kingdom in 1833, it was politically impossible for Great Britain, empty cotton warehouses or not, to recognize, diplomatically, the Confederate States. In addition, during the two years it took to draw down the stockpiles, Britain expanded cotton imports from India, Egypt, and Brazil.

Monopoly sellers often see no threats to their superior marketplace position. In these examples did the power of the monopoly blind the decision makers to other possibilities? Perhaps. As a result of their actions, this is how history unfolded.
MODULE 10: MONOPOLISTIC COMPETITION AND OLIGOPOLY

WHY IT MATTERS: MONOPOLISTIC COMPETITION AND OLIGOPOLY

Why analyze a firm’s profit-maximizing strategies under conditions of monopolistic competition and oligopoly?

The types of firms we’ve covered so far—perfect competition and monopoly—are at opposite ends of the competition spectrum. A perfectly competitive market has many firms selling identical products, who all act as price takers in the face of the competition. If you recall, price takers are firms that have no market power. They simply have to take the market price as given.

Monopoly arises when a single firm sells a product for which there are no close substitutes. We consider Microsoft, for instance, as a monopoly because it dominates the operating systems market.

What about the vast majority of real world firms and organizations that fall between these extremes, firms that we could describe as imperfectly competitive? What determines their behavior? They have more influence over the price they charge than perfectly competitive firms, but not as much as a monopoly. What will they do? For example, consider these following questions:

- Why do gas stations charge different prices for a gallon of gasoline?
- What determines how far apart the prices of Colgate and Crest toothpaste can be?
- Why did fast food restaurants start offering salads?
- Why did McDonalds come up with the Big Mac sandwich?

One type of imperfectly competitive market is monopolistic competition. Monopolistically competitive markets feature a large number of competing firms, but the products that they sell are not identical. Consider, as an example, the Mall of America in Minnesota, the largest shopping mall in the United States. In 2010, the Mall of America had 24 stores that sold women’s “ready-to-wear” clothing (like Ann Taylor and Urban Outfitters), another 50 stores that sold clothing for both men and women (like Banana Republic, J. Crew, and Nordstrom’s), plus 14 more stores that sold women’s specialty clothing (like Motherhood Maternity and Victoria’s Secret). Most of the markets that consumers encounter at the retail level are monopolistically competitive.

The other type of imperfectly competitive market is oligopoly. Oligopolistic markets are those which a small number of firms dominate. Commercial aircraft provides a good example: Boeing and Airbus each produce slightly less than 50% of the large commercial aircraft in the world.
Another example is the U.S. soft drink industry, which Coca-Cola and Pepsi dominate. We characterize oligopolies by high barriers to entry with firms choosing output, pricing, and other decisions strategically based on the decisions of the other firms in the market.

In this module, we first explore how monopolistically competitive firms will choose their profit-maximizing level of output. We will then discuss oligopolistic firms, which face two conflicting temptations: to collaborate as if they were a single monopoly, or to individually compete to gain profits by expanding output levels and cutting prices. Oligopolistic markets and firms can also take on elements of monopoly and of more competitive market models.

In a real sense, the models of monopolistic competition and oligopoly are combinations of the models of perfect competition and monopoly. As you progress through this module, think about the similarities and the differences between each of these models of market structure.

**INTRODUCTION TO MONOPOLISTICALLY COMPETITIVE INDUSTRIES**

What you’ll learn to do: explain the characteristics of monopolistic competition and how it differs from other market structures
Monopolistically competitive industries are those that contain more than a few firms, each of which offers a similar but not identical product. Take fast food, for example. The fast food market is quite competitive, and yet each firm has a monopoly in its own product. Some customers have a preference for McDonald's over Burger King. Some have a preference for Dominoes over Pizza Hut. These preferences give monopolistically competitive firms market power, which they can exploit to earn positive economic profits.

Consider the following questions:

- Why do gas stations charge different prices for a gallon of gasoline?
- What determines how far apart the prices of Colgate and Crest toothpaste can be?
- Why did fast food restaurants start offering salads?
- Why are fast food chicken sandwich prices different from burger prices?
- Why did McDonalds come up with the Big Mac sandwich?
Monopolistic competition is what economists call industries that consist of many firms competing against each other, but selling products that are distinctive in some way. Examples include stores that sell different styles of clothing; restaurants or grocery stores that sell different kinds of food; and even products like golf balls or beer that may be at least somewhat similar but differ in public perception because of advertising and brand names. When products are distinctive, each firm has a mini-monopoly on its particular style or flavor or brand name. However, firms producing such products must also compete with other styles and flavors and brand names. The term “monopolistic competition” captures this mixture of mini-monopoly and tough competition.

Who invented the theory of imperfect competition?

The theory of imperfect competition was developed by two economists independently but simultaneously in 1933. The first was Edward Chamberlin of Harvard University who published The Economics of Monopolistic Competition. The second was Joan Robinson of Cambridge University who published The Economics of Imperfect Competition. Robinson subsequently became interested in macroeconomics where she became a prominent Keynesian, and later a post-Keynesian economist.

Differentiated Products

A firm can try to make its products different from those of its competitors in several ways: physical characteristics of the product, location from which the product is sold, intangible aspects of the product, and perceptions of the product. Products that are distinctive in these ways are called differentiated products.

Physical characteristics of a product include all the phrases you hear in advertisements: unbreakable bottle, nonstick surface, freezer-to-microwave, non-shrink, extra spicy, newly redesigned for your comfort. The location of a firm can also create a difference between producers. For example, a gas station located at a heavily traveled intersection is more convenient than one on a less-traveled back road. A supplier to an automobile manufacturer may find that it is an advantage to locate close to the car factory.

Intangible aspects can differentiate a product, too. Some intangible aspects may be promises like a guarantee of satisfaction or money back, a reputation for high quality, services like free delivery, or offering a loan to purchase the product. Finally, product differentiation may occur in the minds of buyers. For example, many people could not tell the difference in taste between common varieties of beer or cigarettes if they were blindfolded but, because of past habits and advertising, they have strong preferences for certain brands. Advertising can play a role in shaping these intangible preferences.

The concept of differentiated products is closely related to the degree of variety that is available. If everyone in the economy wore only blue jeans, ate only white bread, and drank only tap water, then the markets for clothing, food, and drink would be much closer to perfectly competitive. The variety of styles, flavors, locations, and characteristics creates product differentiation and monopolistic competition.

Perceived Demand for a Monopolistic Competitor

A monopolistically competitive firm perceives a demand for its goods that is an intermediate case between monopoly and competition. Figure 1 offers a reminder that the demand curve as faced by a perfectly competitive firm is perfectly elastic or flat, because the perfectly competitive firm can sell any quantity it wishes at the prevailing market price. In
contrast, the demand curve, as faced by a monopolist, is the market demand curve, since a monopolist is the only firm in the market, and hence is downward sloping.

The demand curve faced by a perfectly competitive firm is perfectly elastic, meaning it can sell all the output it wishes at the prevailing market price. The demand curve faced by a monopoly is the market demand. It can sell more output only by decreasing the price it charges. The demand curve faced by a monopolistically competitive firm falls in between.

The demand curve as faced by a monopolistic competitor is not flat, but rather downward-sloping, which means that the monopolistic competitor can raise its price without losing all of its customers or lower the price and gain more customers. Since there are substitutes, the demand curve facing a monopolistically competitive firm is more elastic than that of a monopoly where there are no close substitutes. If a monopolist raises its price, some consumers will choose not to purchase its product—but they will then need to buy a completely different product. However, when a monopolistic competitor raises its price, some consumers will choose not to purchase the product at all, but others will choose to buy a similar product from another firm. If a monopolistic competitor raises its price, it will not lose as many customers as would a perfectly competitive firm, but it will lose more customers than would a monopoly that raised its prices.

At a glance, the demand curves faced by a monopoly and by a monopolistic competitor look similar—that is, they both slope down. But the underlying economic meaning of these perceived demand curves is different, because a monopolist faces the market demand curve and a monopolistic competitor does not. Rather, a monopolistically competitive firm’s demand curve is but one of many firms that make up the “before” market demand curve. Are you following? If so, how would you categorize the market for golf balls?

**Try It**

Visit this page in your course online to view this presentation.

**ARE GOLF BALLS REALLY DIFFERENTIATED PRODUCTS?**

Monopolistic competition refers to an industry that has more than a few firms, each offering a product which, from the consumer’s perspective, is different from its competitors. The U.S. Golf Association runs a laboratory that tests 20,000 golf balls a year. There are strict rules for what makes a golf ball legal. The weight of a golf ball cannot exceed 1.620 ounces and its diameter cannot be less than 1.680 inches (which is a weight of 45.93 grams and a diameter of 42.67 millimeters, in case you were wondering). The balls are also tested by being hit at different speeds. For example, the distance test involves having a mechanical golfer hit the ball with a titanium driver and a swing speed of 120 miles per hour. As the testing center explains: “The USGA system then uses an array of sensors that accurately measure the flight of a golf ball during a short, indoor trajectory from a ball launcher. From this flight data, a computer
differentiated product: a product that is consumers perceive as distinctive in some way
imperfectly competitive: firms and organizations that fall between the extremes of monopoly and perfect competition
monopolistic competition: many firms competing to sell similar but differentiated products
product differentiation: any action that firms do to make consumers think their products are different from their competitors’

Watch It

Watch this video for a brief overview of monopolistic competition and to see a comparison between perfect competition, monopolistic competition, and monopolies. The video goes on to explain the cost curves for a monopolistically competitive firm and how it compares to those in different competitive settings, which we'll cover in more detail later in the module.

Watch this video online: https://youtu.be/T3F1Vt3lyNc

INTRODUCTION TO ANALYZING AND GRAPHING MONOPOLISTIC COMPETITION

What you’ll learn to do: calculate and graph a firm’s profit in monopolistic competition
In this section, you will learn how to analyze the cost and revenue curves related to monopolistically competitive firms and use these graphs to determine the best price and quantity for a firm's product.

## PROFIT MAXIMIZATION UNDER MONOPOLISTIC COMPETITION

### Learning Objectives
- Describe how a monopolistic competitor chooses price and quantity using marginal revenue and marginal cost
- Graph and interpret a monopolistically competitive firm's average, marginal, and total cost curves
- Compute total revenue, profits, and losses for monopolistic competitors using the demand and average cost curves

### Choosing the Profit-Maximizing Output and Price

The monopolistically competitive firm decides on its profit-maximizing quantity and price in much the same way as a monopolist. A monopolistic competitor, like a monopolist, faces a downward-sloping demand curve, and so it will choose some combination of price and quantity along its perceived demand curve.

As an example of a profit-maximizing monopolistic competitor, consider the Authentic Chinese Pizza store, which serves pizza with cheese, sweet and sour sauce, and your choice of vegetables and meats. Although Authentic Chinese Pizza must compete against other pizza businesses and restaurants, it has a differentiated product. The firm’s perceived demand curve is downward sloping, as shown in Figure 1 and the first two columns of Table 1.
Figure 1. How a Monopolistic Competitor Chooses its Profit Maximizing Output and Price. To maximize profits, the Authentic Chinese Pizza shop would choose a quantity where marginal revenue equals marginal cost, or $Q$ where $MR = MC$. Here it would choose a quantity of 40 and a price of $16.

### Table 1. Revenue and Cost Schedule

<table>
<thead>
<tr>
<th>Quantity</th>
<th>Price</th>
<th>Total Revenue</th>
<th>Marginal Revenue</th>
<th>Total Cost</th>
<th>Marginal Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>$23</td>
<td>$230</td>
<td>—</td>
<td>$340</td>
<td>—</td>
</tr>
<tr>
<td>20</td>
<td>$20</td>
<td>$400</td>
<td>$17</td>
<td>$400</td>
<td>$6</td>
</tr>
<tr>
<td>30</td>
<td>$18</td>
<td>$540</td>
<td>$14</td>
<td>$480</td>
<td>$8</td>
</tr>
<tr>
<td>40</td>
<td>$16</td>
<td>$640</td>
<td>$10</td>
<td>$580</td>
<td>$10</td>
</tr>
<tr>
<td>50</td>
<td>$14</td>
<td>$700</td>
<td>$6</td>
<td>$700</td>
<td>$12</td>
</tr>
<tr>
<td>60</td>
<td>$12</td>
<td>$720</td>
<td>$2</td>
<td>$840</td>
<td>$14</td>
</tr>
<tr>
<td>70</td>
<td>$10</td>
<td>$700</td>
<td>$2</td>
<td>$1,020</td>
<td>$18</td>
</tr>
<tr>
<td>80</td>
<td>$8</td>
<td>$640</td>
<td>$2</td>
<td>$1,280</td>
<td>$26</td>
</tr>
</tbody>
</table>

The combinations of price and quantity at each point on the demand curve can be multiplied to calculate the total revenue that the firm would receive, which is shown in the third column of Table 1. The fourth column, marginal revenue, is calculated as the change in total revenue divided by the change in quantity. The final columns of Table 1 show total cost, marginal cost, and average cost. As always, marginal cost is calculated by dividing the change in total cost by the change in quantity, while average cost is calculated by dividing total cost by quantity. The following example shows how these firms calculate how much of its product to supply at what price.

### How a Monopolistic Competitor Determines How Much to Produce and at What Price

The process by which a monopolistic competitor chooses its profit-maximizing quantity and price resembles closely how a monopoly makes these decisions process. First, the firm selects the profit-maximizing quantity to produce. Then the firm decides what price to charge for that quantity.

**Step 1.** The monopolistic competitor determines its profit-maximizing level of output. In this case, the Authentic Chinese Pizza company will determine the profit-maximizing quantity to produce by considering its marginal revenues and marginal costs. Two scenarios are possible:
If the firm is producing at a quantity of output where marginal revenue exceeds marginal cost, then the firm should keep expanding production, because each marginal unit is adding to profit by bringing in more revenue than its cost. In this way, the firm will produce up to the quantity where MR = MC.

If the firm is producing at a quantity where marginal costs exceed marginal revenue, then each marginal unit is costing more than the revenue it brings in, and the firm will increase its profits by reducing the quantity of output until MR = MC.

In this example, MR and MC intersect at a quantity of 40, which is the profit-maximizing level of output for the firm.

**Step 2.** The monopolistic competitor decides what price to charge. When the firm has determined its profit-maximizing quantity of output, it can then look to its perceived demand curve to find out what it can charge for that quantity of output. On the graph, this process can be shown as a vertical line reaching up through the profit-maximizing quantity until it hits the firm’s perceived demand curve. For Authentic Chinese Pizza, it should charge a price of $16 per pizza for a quantity of 40.

Although the process by which a monopolistic competitor makes decisions about quantity and price is similar to the way in which a monopolist makes such decisions, two differences are worth remembering. First, although both a monopolist and a monopolistic competitor face downward-sloping demand curves, the monopolist’s perceived demand curve is the market demand curve, while the perceived demand curve for a monopolistic competitor is based on the extent of its product differentiation and how many competitors it faces. Second, a monopolist is surrounded by barriers to entry and need not fear entry, but a monopolistic competitor who earns profits must expect the entry of firms with similar, but differentiated, products.

**Try It**

Visit this page in your course online to view this presentation.

**Calculating Profits**

Once the monopolistic competitor has determined the profit-maximizing quantity of output to supply, the next step is to calculate how much profit it is earning. We use the same process that was used with perfect competition and monopoly. This is illustrated in Figure 2, using the data from Table 2, which extends the data from to Table 1 to include average total cost in the last column.

![Figure 2. Computing Profit for a Monopolistic Competitor. To calculate profit, start from the profit-maximizing quantity, which is 40. Next find total revenue which is the area of the rectangle with the height of P = $16 times the base of Q = 40. Next find total cost which is the area of the rectangle with the height of AC = $14.50 times the base of Q = 40. The difference between the two areas is profit, the small rectangle above total cost in the figure.](image-url)
Table 2. Revenue and Cost Schedule, including Average Cost

<table>
<thead>
<tr>
<th>Quantity</th>
<th>Price</th>
<th>Total Revenue</th>
<th>Marginal Revenue</th>
<th>Total Cost</th>
<th>Marginal Cost</th>
<th>Average Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>$23</td>
<td>$230</td>
<td>—</td>
<td>$340</td>
<td>—</td>
<td>$34</td>
</tr>
<tr>
<td>20</td>
<td>$20</td>
<td>$400</td>
<td>$17</td>
<td>$400</td>
<td>$6</td>
<td>$20</td>
</tr>
<tr>
<td>30</td>
<td>$18</td>
<td>$540</td>
<td>$14</td>
<td>$480</td>
<td>$8</td>
<td>$16</td>
</tr>
<tr>
<td>40</td>
<td>$16</td>
<td>$640</td>
<td>$10</td>
<td>$580</td>
<td>$10</td>
<td>$14.50</td>
</tr>
<tr>
<td>50</td>
<td>$14</td>
<td>$700</td>
<td>$6</td>
<td>$700</td>
<td>$12</td>
<td>$14</td>
</tr>
<tr>
<td>60</td>
<td>$12</td>
<td>$720</td>
<td>$2</td>
<td>$840</td>
<td>$14</td>
<td>$14</td>
</tr>
<tr>
<td>70</td>
<td>$10</td>
<td>$700</td>
<td>−$2</td>
<td>$1,020</td>
<td>$18</td>
<td>$14.57</td>
</tr>
<tr>
<td>80</td>
<td>$8</td>
<td>$640</td>
<td>−$6</td>
<td>$1,280</td>
<td>$26</td>
<td>$16</td>
</tr>
</tbody>
</table>

We start by identifying the profit-maximizing level of output, where marginal revenue equals marginal cost. This is $Q = 40$. Next, look for the profit margin, the difference between price and average cost. The price is $16, which you can read off the demand curve for quantity equals 40. The average cost is $14.50, which you can read off the average cost curve for quantity equals 40. The profit margin is $16.00 − $14.50 = $1.50 for each unit that the firm sells. Total profit is the profit margin times the quantity or $1.50 \times 40 = $60$.

Alternatively, we can compute profit as total revenue minus total cost. Total revenue is price times quantity or $16.00 \times 40 = $640$. This is the area of the rectangle that starts at the origin, goes up to a price of $16$, goes over to the demand curve, down to the quantity of 40 and back to the origin. Total cost is average cost times quantity or $14.50 \times 40 = $580$. This is the area of the rectangle that starts at the origin, goes up the vertical axis to an average cost of $14.50$, goes over to the average cost curve, down to the quantity of 40 and back to the origin. Profit is the difference between the two areas, $640 − 580 = 60$. This is shown graphically as the area of the rectangle on top of total cost, or the price minus average cost, times quantity. Note that if the firm was earning zero economic profits, the rectangles of total revenue and total cost would be the same—there would be no profit rectangle. The break-even point occurs where the demand curve intersects with average cost, so $P = AC$. Note also that if the firm was making a loss, the negative profit (i.e. loss) would be the rectangle on top of total revenue.

**Try It**

Visit this page in your course online to check your understanding.

**LEARN BY DOING: PROFIT MAXIMIZATION UNDER MONOPOLISTIC COMPETITION**

**Try It**
ENTRY, EXIT AND PROFITS IN THE LONG RUN

Learning Objectives

- Explain how short run and long run equilibrium affect entry and exit in a monopolistically competitive industry

Monopolistic Competitors and Entry

A monopolistic competitor, like firms in other market structures, may earn profits in the short run, but that doesn't mean they'll be able to keep them. If one monopolistic competitor earns positive economic profits, other firms will be tempted to enter the market. A gas station with a great location must worry that other gas stations might open across the street or down the road—and perhaps the new gas stations will sell coffee or have a carwash or some other attraction to lure customers. A successful restaurant with a unique barbecue sauce must be concerned that other restaurants will try to copy the sauce or offer their own unique recipes. A laundry detergent with a great reputation for quality must be concerned that other competitors may seek to build their own reputations.

The entry of other firms into the same general market (like gas, restaurants, or detergent) shifts the demand curve faced by a monopolistically competitive firm. As more firms enter the market, the quantity demanded at a given price for any particular firm will decline, and the firm's perceived demand curve will shift to the left. As a firm's perceived demand curve shifts to the left, its marginal revenue curve will shift to the left, too. The shift in marginal revenue will change the profit-maximizing quantity that the firm chooses to produce, since marginal revenue will then equal marginal cost at a lower quantity.

Figure 1(a) shows a situation in which a monopolistic competitor was earning a profit with its original perceived demand curve ($D_0$). The intersection of the marginal revenue curve ($MR_0$) and marginal cost curve ($MC$) occurs at point $S$, corresponding to quantity $Q_0$, which is associated on the demand curve at point $T$ with price $P_0$. The combination of
price $P_0$ and quantity $Q_0$ lies above the average cost curve, which shows that the firm is earning positive economic profits.

Figure 1. Monopolistic Competition, Entry, and Exit. (a) At $P_0$ and $Q_0$, the monopolistically competitive firm in this figure is making a positive economic profit. This is clear because if you follow the dotted line above $Q_0$, you can see that price is above average cost. Positive economic profits attract competing firms to the industry, driving the original firm’s demand down to $D_1$. At the new equilibrium quantity ($P_1$, $Q_1$), the original firm is earning zero economic profits, and entry into the industry ceases. In (b) the opposite occurs. At $P_0$ and $Q_0$, the firm is losing money. If you follow the dotted line above $Q_0$, you can see that average cost is above price. Losses induce firms to leave the industry. When they do, demand for the original firm rises to $D_1$, where once again the firm is earning zero economic profit.

Unlike a monopoly, with its high barriers to entry, a monopolistically competitive firm with positive economic profits will attract competition. When another competitor enters the market, the original firm’s perceived demand curve shifts to the left, from $D_0$ to $D_1$, and the associated marginal revenue curve shifts from $MR_0$ to $MR_1$ (as shown in figure 1a). The new profit-maximizing output is $Q_1$, because the intersection of the $MR_1$ and $MC$ now occurs at point $U$. Moving vertically up from that quantity on the new demand curve, the optimal price is at $P_1$.

As long as the firm is earning positive economic profits, new competitors will continue to enter the market, reducing the original firm’s demand and marginal revenue curves. The long-run equilibrium is shown in the figure at point $V$, where the firm’s perceived demand curve touches the average cost curve. When price is equal to average cost, economic profits are zero. Thus, although a monopolistically competitive firm may earn positive economic profits in the short term, the process of new entry will drive down economic profits to zero in the long run. Remember that zero economic profit is not equivalent to zero accounting profit. A zero economic profit means the firm’s accounting profit is equal to what its resources could earn in their next best use. Figure 1(b) shows the reverse situation, where a monopolistically competitive firm is originally losing money. The adjustment to long-run equilibrium is analogous to the previous example. The economic losses lead to firms exiting, which will result in increased demand for this particular firm, and consequently lower losses. Firms exit up to the point where there are no more losses in this market, for example when the demand curve touches the average cost curve, as in point $Z$.

Monopolistic competitors can make an economic profit or loss in the short run, but in the long run, entry and exit will drive these firms toward a zero economic profit outcome. However, the zero economic profit outcome in monopolistic competition looks different from the zero economic profit outcome in perfect competition in several ways relating both to efficiency and to variety in the market.

Try It

Visit this page in your course online to check your understanding.

Watch It
This video demonstrates the graph for a monopolistic competitive firm. In the short run, the graph looks like just like the graph for a monopoly, with the firm making an economic profit. In the long run, however, firms will enter the industry and cause the demand curve to shift to the left, which results in no economic profit.

Watch this video online: https://youtu.be/8a3gXThQeK0

Try It

These questions allow you to get as much practice as you need, as you can click the link at the top of the first question (“Try another version of these questions”) to get a new set of questions. Practice until you feel comfortable doing the questions.

Visit this page in your course online to practice before taking the quiz.

MONOPOLISTIC COMPETITION AND EFFICIENCY

Learning Objectives

- Describe why monopolistically competitive markets are inefficient

Monopolistic Competition and Efficiency

The long-term result of entry and exit in a perfectly competitive market is that all firms end up selling at the price level determined by the lowest point on the average cost curve. This outcome is why perfect competition displays productive efficiency: goods are being produced at the lowest possible average cost. However, in monopolistic competition, the end result of entry and exit is that firms end up with a price that lies on the downward-sloping portion of the average cost curve, not at the very bottom of the AC curve. Thus, monopolistic competition will not be productively efficient.

In a perfectly competitive market, each firm produces at a quantity where price is set equal to marginal cost, both in the short run and in the long run. This outcome is why perfect competition displays allocative efficiency: the social benefits of additional production, as measured by the marginal benefit, which is the same as the price, equal the marginal costs to society of that production. In a monopolistically competitive market, the rule for maximizing profit is to set MR = MC—and price is higher than marginal revenue, not equal to it because the demand curve is downward sloping. When P > MC, which is the outcome in a monopolistically competitive market, the benefits to society of providing additional quantity, as measured by the price that people are willing to pay, exceed the marginal costs to society of producing those units. A monopolistically competitive firm does not produce more, which means that society loses the net benefit of those extra units. This is the same argument we made about monopoly, but in this case to a lesser degree. Thus, a monopolistically competitive industry will produce a lower quantity of a good and charge a higher price for it than would a perfectly competitive industry.

Why Does a Shift in Perceived Demand Cause a Shift in Marginal Revenue?
The combinations of price and quantity at each point on a firm’s perceived demand curve are used to calculate total revenue for each combination of price and quantity. This information on total revenue is then used to calculate marginal revenue, which is the change in total revenue divided by the change in quantity. A change in perceived demand will change total revenue at every quantity of output and in turn, the change in total revenue will shift marginal revenue at each quantity of output. Thus, when entry occurs in a monopolistically competitive industry, the perceived demand curve for each firm will shift to the left, because a smaller quantity will be demanded at any given price. Another way of interpreting this shift in demand is to notice that, for each quantity sold, a lower price will be charged. Consequently, the marginal revenue will be lower for each quantity sold—and the marginal revenue curve will shift to the left as well. Conversely, exit causes the perceived demand curve for a monopolistically competitive firm to shift to the right and the corresponding marginal revenue curve to shift right, too.

A monopolistically competitive industry does not display productive and allocative efficiency in either the short run, when firms are making economic profits and losses, nor in the long run, when firms are earning zero profits.

Try It
Visit this page in your course online to check your understanding.

PRODUCT DIFFERENTIATION AND ADVERTISING

Learning Objectives

- Describe product differentiation and how advertising works in monopolistically competitive industries

The Benefits of Variety and Product Differentiation

Even though monopolistic competition does not provide productive efficiency or allocative efficiency, it does have benefits of its own. Product differentiation is based on variety and innovation. Many people would prefer to live in an economy with many kinds of clothes, foods, and car styles; not in a world of perfect competition where everyone will always wear blue jeans and white shirts, eat only spaghetti with plain red sauce, and drive an identical model of car. Many people would prefer to live in an economy where firms are struggling to figure out ways of attracting customers by methods like friendlier service, free delivery, guarantees of quality, variations on existing products, and a better shopping experience.

Economists have struggled, with only partial success, to address the question of whether a market-oriented economy produces the optimal amount of variety. Critics of market-oriented economies argue that society does not really need dozens of different athletic shoes or breakfast cereals or automobiles. They argue that much of the cost of creating such a high degree of product differentiation, and then of advertising and marketing this differentiation, is socially wasteful—that is, most people would be just as happy with a smaller range of differentiated products produced and sold at a lower price. Defenders of a market-oriented economy respond that if people do not want to buy differentiated products or highly advertised brand names, no one is forcing them to do so. Moreover, they argue that consumers benefit substantially when firms seek short-term profits by providing differentiated products. This controversy may never be fully
resolved, in part because deciding on the optimal amount of variety is very difficult, and in part because the two sides often place different values on what variety means for consumers.

How does advertising impact monopolistic competition?

The U.S. economy spent about $139.5 billion on advertising in 2012, according to Kantar Media Reports. Roughly one third of this was television advertising, and another third was divided roughly equally between Internet, newspapers, and radio. The remaining third was divided up between direct mail, magazines, telephone directory yellow pages, billboards, and other miscellaneous sources. More than 500,000 workers held jobs in the advertising industry.

Advertising is all about explaining to people, or making people believe, that the products of one firm are differentiated from the products of another firm. In the framework of monopolistic competition, there are two ways to conceive of how advertising works: either advertising causes a firm’s perceived demand curve to become more inelastic (that is, it causes the perceived demand curve to become steeper); or advertising causes demand for the firm’s product to increase (that is, it causes the firm’s perceived demand curve to shift to the right). In either case, a successful advertising campaign may allow a firm to sell either a greater quantity or to charge a higher price, or both, and thus increase its profits.

However, economists and business owners have also long suspected that much of the advertising may only offset other advertising. Economist A. C. Pigou wrote the following back in 1920 in his book, *The Economics of Welfare*:

> It may happen that expenditures on advertisement made by competing monopolists [that is, what we now call monopolistic competitors] will simply neutralise one another, and leave the industrial position exactly as it would have been if neither had expended anything. For, clearly, if each of two rivals makes equal efforts to attract the favour of the public away from the other, the total result is the same as it would have been if neither had made any effort at all.

Watch It

Watch this video to see an example of a successful advertising campaign run by the company Axe.

Watch this video online: https://youtu.be/8QNZR0JJfOY

Try It

Visit this page in your course online to check your understanding.

Licensing & Attributions

CC licensed content, Shared previously

- Monopolistic Competition. Authored by: OpenStax College. Located at: https://cnx.org/contents/vEmOH-... License: CC BY: Attribution
- License Terms: Download for free at http://cnx.org/contents/4a84.4:0_gKxtID8@6/Monopolistic-Competition. License: CC BY: Attribution
- Video: Dollars and Scents. Provided by: BBC. Located at: https://www.youtube.com/watch?v=8QNZR0JJfOY. License: CC BY-NC-ND: Attribution-NonCommercial-NoDerivatives
INTRODUCTION TO OLIGOPOLIES

What you’ll learn to do: describe and analyze oligopolies

Most of the firms that get talked about as “monopolies” today or that regulatory authorities pursue antitrust activities against are actually oligopolies, firms that have only a limited number of competitors. There are quite a few industries in the U.S. that are oligopolistic. Think about rental cars, or car manufacturers, or newspapers, or internet service providers.

In this section, you will learn what oligopolies are and why they exist. You’ll read about how some oligopolies are motivated to work together and collude to ensure higher profits, while others compete and act more like perfect competitors.

WHY DO OLIGOPOLIES EXIST?

Learning Objectives

- Describe why and how oligopolies exist
Laundry detergent and bags of ice—products of industries that seem pretty mundane, maybe even boring. Hardly! Both have been the center of clandestine meetings and secret deals worthy of a spy novel. In France, between 1997 and 2004, the top four laundry detergent producers (Proctor & Gamble, Henkel, Unilever, and Colgate-Palmolive) controlled about 90 percent of the French soap market. Officials from the soap firms were meeting secretly, in out-of-the-way, small cafés around Paris. Their goals: stamp out competition and set prices.

Around the same time, the top five Midwest ice makers (Home City Ice, Lang Ice, Tinley Ice, Sisler’s Dairy, and Products of Ohio) had similar goals in mind when they secretly agreed to divide up the bagged ice market. If both groups could meet their goals, it would enable each to act as though they were a single firm—in essence, a monopoly—and enjoy monopoly-size profits. The problem? In many parts of the world, including the European Union and the United States, it is illegal for firms to divide up markets and set prices collaboratively.

These two cases provide examples of markets that are characterized neither as perfect competition nor monopoly. Instead, these firms are competing in market structures that lie between the extremes of monopoly and perfect competition. How do they behave? Why do they exist?

Oligopolistic markets are those dominated by a small number of firms. Think of the U.S. soft drink industry, which is dominated by Coca-Cola and Pepsi. Oligopolies are characterized by high barriers to entry with firms strategically choosing output, pricing, and other decisions based on the decisions of the other firms in the market.

Why Do Oligopolies Exist?

Many purchases that individuals make at the retail level are produced in markets that are neither perfectly competitive, monopolies, nor monopolistically competitive. Rather, they are oligopolies. Oligopoly arises when a small number of large firms have all or most of the sales in an industry. Examples of oligopoly abound and include the auto industry, cable television, and commercial air travel. Oligopolistic firms are like cats in a bag. They can either scratch each other to pieces or cuddle up and get comfortable with one another. If oligopolists compete hard, they may end up acting very much like perfect competitors, driving down costs and leading to zero profits for all. If oligopolists collude with each other, they may effectively act like a monopoly and succeed in pushing up prices and earning consistently high levels of profit. Oligopolies are typically characterized by mutual interdependence where various decisions such as output, price, advertising, and so on, depend on the decisions of the other firm(s). Analyzing the choices of oligopolistic firms about pricing and quantity produced involves considering the pros and cons of competition versus collusion at a given point in time.
A combination of the barriers to entry that create monopolies and the product differentiation that characterizes monopolistic competition can create the setting for an oligopoly. For example, when a government grants a patent for an invention to one firm, it may create a monopoly. When the government grants patents to, for example, three different pharmaceutical companies that each has its own drug for reducing high blood pressure, those three firms may become an oligopoly.

Similarly, a natural monopoly will arise when the demand in a market is only large enough for a single firm to operate at the minimum of the long-run average cost curve. In such a setting, the market has room for only one firm, because no smaller firm can operate at a low enough average cost to compete, and no larger firm could sell what it produced given the quantity demanded in the market.

The demand may also only be limited to two or three times the quantity needed to produce at the minimum of the average cost curve—which means that the market would have room for only two or three oligopoly firms (and they need not produce differentiated products). Again, smaller firms would have higher average costs and be unable to compete, while additional large firms would produce such a high quantity that they would not be able to sell it at a profitable price. This combination of economies of scale and market demand creates the barrier to entry, which led to the Boeing-Airbus oligopoly for large passenger aircraft.

The product differentiation at the heart of monopolistic competition can also play a role in creating oligopoly. For example, firms may need to reach a certain minimum size before they are able to spend enough on advertising and marketing to create a recognizable brand name. The problem in competing with, say, Coca-Cola or Pepsi is not that producing fizzy drinks is technologically difficult, but rather that creating a brand name and marketing effort to equal Coke or Pepsi is an enormous task.

Try It
Visit this page in your course online to check your understanding.

Oligopoly Models

There is no single, generally accepted model of oligopoly, but rather there are a number of models that apply in different situations. We will discuss several of these in the following pages. In principle, one can calculate and graph an oligopoly’s cost and revenue curves, and determine its profit maximizing level of output and price in the same way as we did with monopoly. What complicates matters with oligopolistic industries is that any one firm’s demand and marginal revenue curves are influenced by what the other oligopolistic firms are doing. For example, if Pepsi goes on sale, the demand for Coca-Cola declines, with the demand and marginal revenue curves shifting to the left. Thus, the best price and quantity for any oligopoly depends on what every other firm in the industry is doing.
Explain collusion and cartels
Explain and analyze profits and losses in an oligopoly (including how to maximize profits and how cut-throat competition can result in zero economic profits)
Explain why oligopolies are inefficient

When oligopolist firms consider what quantity to produce and what price to charge, they face a temptation to work with the other firms to act as if they were a single monopoly. By acting together, oligoplistic firms can hold down industry output, charge a higher price, and divide the profit among themselves. When firms act together in this way to reduce output and keep prices high, it is called collusion. A group of firms that have a formal agreement to collude to produce the monopoly output and sell at the monopoly price is called a cartel.

We can see what collusion looks like in Figure 1. If the firms decide to collude, they choose to produce the monopoly output, Qc, and charge a corresponding price, Pc, which can be read off the market demand curve. Since they produce together where MR = MC, they will maximize industry profits, just like an actual monopoly would.

![Figure 1. Profit Maximization for an Oligopoly.](image)

**Figure 1. Profit Maximization for an Oligopoly.** The profit maximizing point for colluding oligopolies is found where MR=MC, where price is \( P_c \), just as in a monopoly. Because of cutthroat competition, oligopolies may instead act as perfect competitors, moving the profit maximizing point to where demand and MC intersect, just as in perfect competition. This is found at the intersection of \( Q_{cc} \) and \( P_{cc} \).

**COLLUSION VERSUS CARTELS: HOW TO DIFFERENTIATE**

In the United States, as well as many other countries, it is illegal for firms to collude since collusion is anti-competitive behavior, which is a violation of antitrust law. Both the Antitrust Division of the Justice Department and the Federal Trade Commission have responsibilities for preventing collusion in the United States.

The problem of enforcement is finding hard evidence of collusion. Cartels are formal agreements to collude. Because cartel agreements provide evidence of collusion, they are rare in the United States. Instead, most collusion is tacit, where firms implicitly reach an understanding that competition is bad for profits.

Economists have understood for a long time the desire of businesses to avoid competing so that they can instead raise the prices that they charge and earn higher profits. Adam Smith wrote in *Wealth of Nations* in 1776: “People of the same trade seldom meet together, even for merriment and diversion, but the conversation ends in a conspiracy against the public, or in some contrivance to raise prices.”

**Try It**

Visit this page in your course online to view this presentation.
Cut-throat Competition

Even when oligopolists recognize that they would benefit as a group by acting like a monopoly, each individual oligopoly faces a private temptation to produce just a slightly higher quantity and earn slightly higher profit—while still counting on the other oligopolists to hold down their production and keep prices high. If at least some oligopolists give in to this temptation and start producing more, then the market price will fall. A small handful of oligopoly firms may end up competing so fiercely that they all find themselves earning zero economic profits—as if they were perfect competitors. This situation is called cut-throat competition, and is shown in Figure 1 at Qcc and Pcc. Since Pcc equals average cost, firms end up just breaking even.

Watch It

Watch this video for an explanation of collusion and to learn more about why cartels often fall apart.
Watch this video online: https://youtu.be/BYLe3ErmFFA

Try It

Visit this page in your course online to view this presentation.

Tradeoffs of Imperfect Competition

Oligopoly is probably the second most common market structure (monopolistic competition being the first). When oligopolies result from patented innovations or from taking advantage of economies of scale to produce at low average cost, they may provide considerable benefit to consumers. Oligopolies are often protected by significant barriers to entry, which enable the oligopolists to earn sustained profits over long periods of time. They typically operate at a level of output where price is greater than marginal cost, so oligopolistic industries are not allocatively efficient. Unlike in the simple example in Figure 1, oligopolists also do not typically produce at the minimum of their average cost curves, so they are not productively efficient. When they lack vibrant competition, they may lack incentives to provide innovative products and high-quality service.

The task of public policy with regard to competition is to sort through these multiple realities, attempting to encourage behavior that is beneficial to the broader society and to discourage behavior that only adds to the profits of a few large companies, with no corresponding benefit to consumers.

Try It

Visit this page in your course online to check your understanding.

THE TEMPTATION TO DEFY THE LAW

Oligopolistic firms have been called “cats in a bag,” as mentioned earlier. The French detergent makers we mentioned at the beginning of our discussion on oligopolies chose to “cozy up” with each other. The result? An uneasy and tenuous relationship. When the Wall Street Journal reported on the matter, it wrote: “According to a statement a Henkel manager made to the [French anti-trust] commission, the detergent makers wanted ‘to limit the intensity of the competition between them and clean up the market.’ Nevertheless, by the early 1990s, a price war had broken out among them.” During the soap executives’ meetings, which sometimes lasted more than four hours, complex pricing structures were established. “One [soap] executive recalled ‘chaotic’ meetings as each side tried to
work out how the other had bent the rules.” Like many cartels, the soap cartel disintegrated due to the very strong temptation for each member to maximize its own individual profits.

How did this soap opera end? After an investigation, French antitrust authorities fined Colgate-Palmolive, Henkel, and Proctor & Gamble a total of €361 million ($484 million). A similar fate befell the icemakers. Bagged ice is a commodity, a perfect substitute, generally sold in 7- or 22-pound bags. No one cares what label is on the bag. By agreeing to carve up the ice market, control broad geographic swaths of territory, and set prices, the icemakers moved from perfect competition to a monopoly model. After the agreements, each firm was the sole supplier of bagged ice to a region; there were profits in both the long run and the short run. According to the courts: “These companies illegally conspired to manipulate the marketplace.” Fines totaled about $600,000—a steep fine considering a bag of ice sells for under $3 in most parts of the United States.

Even though it is illegal in many parts of the world for firms to set prices and carve up a market, the temptation to earn higher profits makes it extremely tempting to defy the law.

Glossary

**cartel**: a group of firms that collude to produce the monopoly output and sell at the monopoly price

**collusion**: when firms act together to reduce output and keep prices high

**cut-throat competition**: oligopolistic outcome when firms decide to cut prices to capture market share; in the limit, this leads to zero economic profits

---

**PRISONER'S DILEMMA**

Learning Objectives

- Explain the role of game theory in understanding the behavior of oligopolies

Prisoner’s Dilemma

Because of the complexity of oligopoly, which is the result of mutual interdependence among firms, there is no single, generally-accepted theory of how oligopolies behave, in the same way that we have theories for all the other market structures. Instead, many economists use game theory, a branch of mathematics that analyzes situations in which players must make decisions and then receive payoffs based on what other players decide to do. Game theory has found widespread applications in the social sciences, as well as in business, law, and military strategy.

A key element of game theory is the concept of Nash equilibrium. The concept was developed by John Nash, an American mathematician who was awarded the 1994 Nobel Prize in economics for this work. A Nash equilibrium occurs when no player has an incentive to change their decision, taking into account what the players have decided and assuming the other players don’t change their decisions. Thus, all players have made an optimal decision, given the decisions of the other players.
The prisoner's dilemma is a scenario in which the gains from cooperation are larger than the rewards from pursuing self-interest. It applies well to oligopoly. The story behind the prisoner's dilemma goes like this:

Two co-conspiratorial criminals are arrested. When they are taken to the police station, they refuse to say anything and are put in separate interrogation rooms. Eventually, a police officer enters the room where Prisoner A is being held and says: “You know what? Your partner in the other room is confessing. So your partner is going to get a light prison sentence of just one year, and because you're remaining silent, the judge is going to stick you with eight years in prison. Why don’t you get smart? If you confess, too, we'll cut your jail time down to five years, and your partner will get five years, also.” Over in the next room, another police officer is giving exactly the same speech to Prisoner B. What the police officers do not say is that if both prisoners remain silent, the evidence against them is not especially strong, and the prisoners will end up with only two years in jail each.

The game theory situation facing the two prisoners is shown in Table 1. To understand the dilemma, first consider the choices from Prisoner A's point of view. If A believes that B will confess, then A ought to confess, too, so as to not get stuck with the eight years in prison. But if A believes that B will not confess, then A will be tempted to act selfishly and confess, so as to serve only one year. The key point is that A has an incentive to confess regardless of what choice B makes! B faces the same set of choices, and thus will have an incentive to confess regardless of what choice A makes. Confess is considered the dominant strategy or the strategy an individual (or firm) will pursue regardless of the other individual's (or firm's) decision. The result is that if prisoners pursue their own self-interest, both are likely to confess, and end up doing a total of 10 years of jail time between them. You should note that this result is a Nash equilibrium.

The game is called a dilemma because if the two prisoners had cooperated by both remaining silent, they would only have had to serve a total of four years of jail time between them. If the two prisoners can work out some way of cooperating so that neither one will confess, they will both be better off than if they each follow their own individual self-interest, which in this case leads straight into longer jail terms.

The Oligopoly Version of the Prisoner’s Dilemma

The members of an oligopoly can face a prisoner’s dilemma, also. If each of the oligopolists cooperates in holding down output, then high monopoly profits are possible. Each oligopolist, however, must worry that while it is holding down output, other firms are taking advantage of the high price by raising output and earning higher profits. Table 2 shows the prisoner’s dilemma for a two-firm oligopoly—known as a duopoly. If Firms A and B both agree to hold down output, they are acting together as a monopoly and will each earn $1,000 in profits. However, both firms’ dominant strategy is to increase output, in which case each will earn $400 in profits.

<table>
<thead>
<tr>
<th>Table 1. The Prisoner's Dilemma Problem</th>
<th>Prisoner B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Remain Silent (cooperate with other prisoner)</td>
</tr>
<tr>
<td><strong>Prisoner A</strong></td>
<td><strong>A gets 2 years, B gets 2 years</strong></td>
</tr>
<tr>
<td>Remain Silent (cooperate with other prisoner)</td>
<td>A gets 1 year, B gets 8 years</td>
</tr>
<tr>
<td>Confess (do not cooperate with other prisoner)</td>
<td>A gets 1 year, B gets 8 years</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 2. A Prisoner’s Dilemma for Oligopolists</th>
<th>Firm B</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Firm B</strong></td>
<td><strong>Increase Output (do not cooperate with other firm)</strong></td>
</tr>
<tr>
<td><strong>Firm A</strong></td>
<td><strong>A gets $1,500, B gets $200</strong></td>
</tr>
<tr>
<td>Hold Down Output (cooperate with other firm)</td>
<td><strong>A gets $1,000, B gets $1,000</strong></td>
</tr>
<tr>
<td>Increase Output (do not cooperate with other firm)</td>
<td><strong>A gets $1,000, B gets $1,000</strong></td>
</tr>
</tbody>
</table>
Can the two firms trust each other? Consider the situation of Firm A:

- If A thinks that B will cheat on their agreement and increase output, then A will increase output, too, because for A the profit of $400 when both firms increase output (the bottom right-hand choice in Table 2) is better than a profit of only $200 if A keeps output low and B raises output (the upper right-hand choice in the table).
- If A thinks that B will cooperate by holding down output, then A may seize the opportunity to earn higher profits by raising output. After all, if B is going to hold down output, then A can earn $1,500 in profits by expanding output (the bottom left-hand choice in the table) compared with only $1,000 by holding down output as well (the upper left-hand choice in the table).

Thus, firm A will reason that it makes sense to expand output if B holds down output and that it also makes sense to expand output if B raises output. Again, B faces a parallel set of decisions.

The result of this prisoner’s dilemma is often that even though A and B could make the highest combined profits by cooperating in producing a lower level of output and acting like a monopolist, the two firms may well end up in a situation where they each increase output and earn only $400 each in profits. The following example discusses one cartel scandal in particular.

What is the Lysine cartel?

Lysine, a $600 million-a-year industry, is an amino acid used by farmers as a feed additive to ensure the proper growth of swine and poultry. The primary U.S. producer of lysine is Archer Daniels Midland (ADM), but several other large European and Japanese firms are also in this market. For a time in the first half of the 1990s, the world’s major lysine producers met together in hotel conference rooms and decided exactly how much each firm would sell and what it would charge. The U.S. Federal Bureau of Investigation (FBI), however, had learned of the cartel and placed wire taps on a number of their phone calls and meetings.

From FBI surveillance tapes, following is a comment that Terry Wilson, president of the corn processing division at ADM, made to the other lysine producers at a 1994 meeting in Mona, Hawaii:

I wanna go back and I wanna say something very simple. If we’re going to trust each other, okay, and if I’m assured that I’m gonna get 67,000 tons by the year’s end, we’re gonna sell it at the prices we agreed to . . . . The only thing we need to talk about there because we are gonna get manipulated by these [expletive] buyers—they can be smarter than us if we let them be smarter. . . . They [the customers] are not your friend. They are not my friend. And we gotta have ‘em, but they are not my friends. You are my friend. I wanna be closer to you than I am to any customer. Cause you can make us … money. … And all I wanna tell you again is let’s—let’s put the prices on the board. Let’s all agree that’s what we’re gonna do and then walk out of here and do it.

The price of lysine doubled while the cartel was in effect. Confronted by the FBI tapes, Archer Daniels Midland pled guilty in 1996 and paid a fine of $100 million. A number of top executives, both at ADM and other firms, later paid fines of up to $350,000 and were sentenced to 24–30 months in prison.

In another one of the FBI recordings, the president of Archer Daniels Midland told an executive from another competing firm that ADM had a slogan that, in his words, had “penetrated the whole company.” The company president stated the slogan this way: “Our competitors are our friends. Our customers are the enemy.” That slogan could stand as the motto of cartels everywhere.

How to Enforce Cooperation

How can parties who find themselves in a prisoner’s dilemma situation avoid the undesired outcome and cooperate with each other? The way out of a prisoner’s dilemma is to find a way to penalize those who do not cooperate.

Perhaps the easiest approach for colluding oligopolists, as you might imagine, would be to sign a contract with each other that they will hold output low and keep prices high. If a group of U.S. companies signed such a contract, however, it would be illegal. Certain international organizations, like the nations that are members of the Organization of Petroleum Exporting Countries (OPEC), have signed international agreements to act like a monopoly, hold down output, and keep prices high so that all of the countries can make high profits from oil exports. Such agreements, however, because they fall in a gray area of international law, are not legally enforceable. If Nigeria, for example, decides to start cutting prices and selling more oil, Saudi Arabia cannot sue Nigeria in court and force it to stop.
Because oligopolists cannot sign a legally enforceable contract to act like a monopoly, the firms may instead keep close tabs on what other firms are producing and charging. Alternatively, oligopolists may choose to act in a way that generates pressure on each firm to stick to its agreed quantity of output.

One example of the pressure these firms can exert on one another is the kinked demand curve, in which competing oligopoly firms commit to match price cuts, but not price increases. This situation is shown in Figure 1. Say that an oligopoly airline has agreed with the rest of a cartel to provide a quantity of 10,000 seats on the New York to Los Angeles route, at a price of $500. This choice defines the kink in the firm’s perceived demand curve. The reason that the firm faces a kink in its demand curve is because of how the other oligopolists react to changes in the firm’s price. If the oligopoly decides to produce more and cut its price, the other members of the cartel will immediately match any price cuts—and therefore, a lower price brings very little increase in quantity sold.

If one firm cuts its price to $300, it will be able to sell only 11,000 seats. However, if the airline seeks to raise prices, the other oligopolists will not raise their prices, and so the firm that raised prices will lose a considerable share of sales. For example, if the firm raises its price to $550, its sales drop to 5,000 seats sold. Thus, if oligopolists always match price cuts by other firms in the cartel, but do not match price increases, then none of the oligopolists will have a strong incentive to change prices, since the potential gains are minimal. This strategy can work like a silent form of cooperation, in which the cartel successfully manages to hold down output, increase price, and share a monopoly level of profits even without any legally enforceable agreement.

Many real-world oligopolies, prodded by economic changes, legal and political pressures, and the egos of their top executives, go through episodes of cooperation and competition. If oligopolies could sustain cooperation with each other on output and pricing, they could earn profits as if they were a single monopoly. However, each firm in an oligopoly has an incentive to produce more and grab a bigger share of the overall market; when firms start behaving in this way, the market outcome in terms of prices and quantity can be similar to that of a highly competitive market.
Watch It

Watch this video to review the key characteristics of oligopolies and to see some applications of game theory and collusion.
Watch this video online: https://youtu.be/PCcVODWm-oY

Try It

Visit this page in your course online to check your understanding.

Glossary

duopoly: an oligopoly with only two firms

game theory: a branch of mathematics that economists use to analyze situations in which players must make decisions and then receive payoffs based on what decisions the other players make

kinked demand curve: a perceived demand curve that arises when competing oligopoly firms commit to match price cuts, but not price increases

Nash equilibrium: solution to a game-theoretic scenario when no player has an incentive to change their decision, taking into account what the players have decided and assuming the other players don’t change their decisions.

prisoner’s dilemma: a game in which the gains from cooperation are larger than the rewards from pursuing self-interest

LEARN BY DOING: PRISONER'S DILEMMA

Try It

These questions allow you to get as much practice as you need, as you can click the link at the top of the first question (“Try another version of these questions”) to get a new set of questions. Practice until you feel comfortable doing the questions.

Visit this page in your course online to practice before taking the quiz.

Try It
Try It

These questions allow you to get as much practice as you need, as you can click the link at the top of the first question (“Try another version of these questions”) to get a new set of questions. Practice until you feel comfortable doing the questions.
Visit this page in your course online to practice before taking the quiz.

PUTTING IT TOGETHER: MONOPOLISTIC COMPETITION AND OLIGOPOLY

Monopolistically competitive industries consist of a significant number of firms, which each produce a differentiated (or heterogeneous) production. In other words, Colgate, AIM, and Tom’s of Maine each produce toothpaste, but they are not identical products. Like firms in any market structure, if a monopolistically competitive firm wishes to maximize profits, it will supply the quantity of output where marginal revenue equals marginal cost. Like perfectly competitive firms, competition prevents monopolistically competitive firms from earning positive economic profits in the long run, unless those firms create innovative new products and/or use advertising to convince customers they have done so.

Returning to some of the questions posed in the “Why it Matters” feature at the beginning of this module:

- Why do gas stations charge different prices for a gallon of gasoline? Some gasoline companies use different additives to make their products at least appear different. This allows them to charge higher prices than companies that don’t make as good a case for their product. Location also matters. A gas station just off the highway can charge higher prices than stations further away, because travelers perceive and are willing to pay for the convenience of the former.
- What determines how far apart the prices of Colgate and Crest toothpaste can be? The answer is brand name loyalty. To the extent that Colgate users believe Colgate is superior to Crest, they will be willing to pay more for Colgate than for Crest. By contrast, if the two products are perceived to be close substitutes, the prices should be similar.
- Why did fast food restaurants start offering salads? Fast food restaurants added salads to their menus to differentiate their product by appealing to health conscious diners.
Why did McDonalds come up with the Big Mac sandwich? McDonalds invented the Big Mac because its competitors offered similar enough regular burgers that McDonalds lost its monopoly profits. The Big Mac restored those profits, at least until Burger King came up with the Whopper and other fast food restaurants developed their own special burgers.

While oligopoly is defined as an industry consisting of, or dominated by a small number of firms, the key characteristic is interdependence among firms. Oligopolies can be characterized by collusion, where firms act jointly like a monopolist to share industry profits, or by competition, where firms compete aggressively for individual profits, or something in between. The computer operating system, dominated by Microsoft, fits the former profile with persistent high economic profits. The airline industry (e.g. United) fits the latter profile, leading to prices barely above costs and low profits.

Oligopolies are inefficient for the same reasons that monopolies are—in order to reap economic profits, they produce too little output so they create deadweight losses to society. The more like a monopoly a given oligopoly is, the higher their profits and the greater the deadweight loss. This is why strong oligopolies usually generate antitrust action by the government.

Licensing & Attributions

CC licensed content, Original

- Putting It Together: Monopolistic Competition and Oligopoly. Authored by Steven Greenlaw and Lumen Learning. License: CC BY Attribution
Why describe the complications surrounding public goods and externalities in an economy?

Decisions made by firms and individuals in a market often have a spillover effect on other people, whether it be for good (like your neighbor’s sweet-smelling cookies), or for the bad (the smelly paper plant that opened up a mile away). In this module, we learn about these positive and negative externalities, as well as the range of public versus private goods.

One of the characteristics of any good or service is its “public” or “private”-ness. Goods can be public or private, or anything in between. What does this mean? It’s probably not what you think. It doesn’t primarily mean who provides the product. In other words, a good is not private because it is provided by private businesses. A good is not public because it is provided by the government. Rather, private/public is more a description of how these goods are consumed.

All of us consume private goods and public goods. There are three basic cases: A private good is one for which the consumer pays all the costs and receives all the benefits. If you buy an apple to eat, no one gets to share any part of the apple that you eat. A public good is one where one person’s consumption doesn’t prevent anyone else from consuming it too. National defense is the classic example—once it’s there, everyone gets the benefits of it. This leads to “free riding,” where people try to avoid paying for the public good since they’ll get the benefits anyway.

In between public and private goods are externality goods (or semi-public goods). In some cases, benefits go beyond the individuals who consumed the externality good. One example is health expenditures. People in a community at large benefit when others get vaccinated against disease, since people won’t be infected by people who are vaccinated. In other cases, costs are imposed on people other than the ones who produced the externality good. Suppose you go to a park for a picnic, and you bring your iPod and external speakers. You play music during the picnic, which others in the park find objectionable. Your music reduces the quality of other park visitors’ experience.

- What are some other examples of private goods? (Hint: most of the things you consume are private goods.)
- What are some other examples of public goods?
- How can “free riding” be prevented?
- What are some other examples of externality goods (or externalities in general)?
- Why does it make sense for society to pay for public education, when not everyone has children?
Public goods and externality goods tend to be either supplied too much or supplied too little compared to what is optimal for society. What policies can government pursue to ameliorate this?

Let’s find out.

INTRODUCTION TO PUBLIC GOODS AND EXTERNALITIES

What you’ll learn to do: define and give examples of public goods and externalities

Figure 1. Roads are an example of a public good.

We’ve learned that free markets are socially optimal (or more specifically, allocatively efficient) because they provide the quantity of output that maximizes the social surplus. In this section, we will learn about how markets for certain products, i.e. public goods and goods with externalities, can fail to provide the socially optimal quantity of a product.
Learning Objectives

- Define externalities and market failure
- Explain how markets do not always allocate goods efficiently, due to externalities

Externalities

Markets offer an efficient way to put buyers and sellers together to determine the quantity of goods which will be produced, the price that will be charged. The principle that voluntary exchange benefits both buyers and sellers is a fundamental building block of the economic way of thinking. But the efficiency of markets depends on the assumption that only the buyer and seller are affected by the transaction. What happens when a voluntary exchange affects a third party who is neither the buyer nor the seller?

When a market does not operate efficiently, the result is called market failure. Markets usually work best when there are no unintended side effects, but that's not always the case. Sometimes people share in the benefits of others’ production or consumption. For example, when you get the flu shot, your neighbors benefit also by not getting the disease from you. They benefit from a side effect of your consumption. Or, you can be negatively impacted by the decisions of others. If your neighbor doesn’t mow their lawn or maintain their home, that hurts the value of your home. In this situation, laissez faire is not the best policy.

Try It

Visit this page in your course online to check your understanding.

watch It

Watch this video to see how externalities and public goods are examples of market failure.
Watch this video online: https://youtu.be/rJixtB0GluQ

Consider an example of a concert producer who wants to build an outdoor arena that will host country music concerts a half-mile from your neighborhood. You will be able to hear these outdoor concerts while sitting on your back porch—or perhaps even in your dining room. In this case, the sellers and buyers of concert tickets may both be quite satisfied with their voluntary exchange, but you have no voice in their market transaction. The effect of a market exchange on a third party who is outside or “external” to the exchange is called an externality. Because externalities that occur in market transactions affect other parties beyond those involved, they are sometimes called spillovers.

Externalities can be negative or positive. If you hate country music, then having it waft into your house every night would be a negative externality. If you love country music, then what amounts to a series of free concerts would be a positive externality.

THE BENEFITS OF VOYAGER I ENDURE

The rapid growth of technology has increased our ability to access and process data, to navigate through a busy city, and to communicate with friends on the other side of the globe. The research and development efforts of citizens, scientists, firms, universities, and governments have truly revolutionized the modern economy. To get a sense of how
far we have come in a short period of time, let’s compare one of humankind’s greatest achievements to the smartphone most of us have in our coat pocket.

In 1977 the United States launched Voyager I, a spacecraft originally intended to reach Jupiter and Saturn, to send back photographs and other cosmic measurements. Voyager I, however, kept going, and going—past Jupiter and Saturn—right out of our solar system. At the time of its launch, Voyager had some of the most sophisticated computing processing power NASA could engineer (8,000 instructions per second), but today, we Earthlings use handheld devices that can process 14 billion instructions per second.

Still, the technology of today is a spillover product of the incredible feats NASA accomplished forty years ago. NASA research, for instance, is responsible for the kidney dialysis and mammogram machines that we use today. Research in new technologies not only produces private benefits to the investing firm, or in this case to NASA, but it also creates benefits for the broader society. In this way, new knowledge often becomes what economists refer to as a public good. This leads us to the topic of this chapter—technology, positive externalities, public goods, and the role of government in encouraging innovation and the social benefits that it provides.

Positive Externalities in Public Health Programs

One of the most remarkable changes in the standard of living in the last several centuries is that people are living longer. Thousands of years ago, human life expectancy is believed to have been in the range of 20 to 30 years. By 1900, average life expectancy in the United States was 47 years. By the start of the twenty-first century, U.S. life expectancy was 77 years. Most of the gains in life expectancy in the history of the human race happened in the twentieth century.

The rise in life expectancy seems to stem from three primary factors. First, systems for providing clean water and disposing of human waste helped to prevent the transmission of many diseases. Second, changes in public behavior have advanced health. Early in the twentieth century, for example, people learned the importance of boiling bottles before using them for food storage and baby’s milk, washing their hands, and protecting food from flies. More recent behavioral changes include reducing the number of people who smoke tobacco and precautions to limit sexually transmitted diseases. Third, medicine has played a large role. Immunizations for diphtheria, cholera, pertussis, tuberculosis, tetanus, and yellow fever were developed between 1890 and 1930. Penicillin, discovered in 1941, led to a series of other antibiotic drugs for bringing infectious diseases under control. In recent decades, drugs that reduce the risks of high blood pressure have had a dramatic effect in extending lives.

These advances in public health have all been closely linked to positive externalities and public goods. Public health officials taught hygienic practices to mothers in the early 1900s and encouraged less smoking in the late 1900s. Many public sanitation systems and storm sewers were funded by government because they have the key traits of public goods. In the twentieth century, many medical discoveries came out of government or university-funded research. Patents and intellectual property rights provided an additional incentive for private inventors. The reason for requiring immunizations, phrased in economic terms, is that it prevents spillovers of illness to others—as well as helping the person immunized.

Watch It

Watch this video to examine the costs and benefits of both positive and negative externalities. We’ll examine how these externalities affect the market and influence the graph of supply and demand in more detail soon.

Watch this video online: https://youtu.be/yC5R9WPId0s

Try It

Visit this page in your course online to check your understanding.

Glossary
**externality**: a market exchange that affects a third party who is outside or “external” to the exchange; sometimes called a “spillover”

**market failure**: When the market on its own does not allocate resources efficiently in a way that balances social costs and benefits; externalities are one example of a market failure

**negative externality**: a situation where a third party, outside the transaction, suffers from a market transaction by others

**positive externality**: a situation where a third party, outside the transaction, benefits from a market transaction by others

**spillover**: see externality

---

**Learning Objectives**

- Explain characteristics of public goods
- Describe common resources and the tragedy of the commons

Consider a good or service where the positive externalities are so extensive that the majority of the benefits that come from the product are external to the person who purchases it, or the firms who produce it. This kind of good is called a public good. Spending on national defense is a good example of a public good. Let’s begin by defining the characteristics of a public good and discussing why these characteristics make it difficult for private firms to supply public goods. Then we will see how government may step in to address the issue.

**The Definition of a Public Good**

To understand the defining characteristics of a public good, first consider an ordinary private good, like a piece of pizza. A piece of pizza can be bought and sold fairly easily because it is a separate and identifiable item. However, public goods are not separate and identifiable in this way.

Instead, public goods have two defining characteristics: they are nonexcludable and nonrivalrous. The first characteristic, that a public good is **nonexcludable**, means that it is costly or impossible to exclude someone from using
If Larry buys a private good like a piece of pizza, then he can exclude others, like Lorna, from eating that pizza. However, if national defense is being provided, then it includes everyone. Even if you strongly disagree with America's defense policies or with the level of defense spending, the national defense still protects you. You cannot choose to be unprotected, and national defense cannot protect everyone else and exclude you.

The second main characteristic of a public good, that it is **nonrivalrous**, means that when one person uses the public good, another can also use it. With a private good like pizza, if Max is eating the pizza then Michelle cannot also eat it; that is, the two people are rivals in consumption. With a public good like national defense, Max's consumption of national defense does not reduce the amount left for Michelle, so they are nonrivalrous in this area.

A number of government services are examples of public goods. For instance, it would not be easy to provide fire and police service so that some people in a neighborhood would be protected from the burning and burglary of their property, while others would not be protected at all. Protecting some necessarily means protecting others, too.

Positive externalities and public goods are closely related concepts. Public goods like police protection or public health funding, have positive externalities, 1. Not all goods and services with positive externalities, however, are public goods. Investments in education have huge positive spillovers but since the majority of the benefits still are received by the students, education is not a public good. Private companies can invest in new inventions such as the Apple iPad and reap profits that may not capture all of the social benefits. Patents can also be described as an attempt to make new inventions into private goods, which are excludable and rivalrous, so that no one but the inventor is allowed to use them during the length of the patent.

Try It

Visit this page in your course online to check your understanding.

Watch It

Watch this video to review the definitions of nonexcludable and nonrivalrous and to learn about more examples of public goods.

Watch this video online: [https://youtu.be/mYtW1Ug7L_s](https://youtu.be/mYtW1Ug7L_s)

**Common Resources and the “Tragedy of the Commons”**

There are some goods that do not fall neatly into the categories of private good or public good. While it is easy to classify a pizza as a private good and a city park as a public good, what about an item that is nonexcludable and rivalrous, such as the queen conch?

In the Caribbean, the queen conch is a large marine mollusk found in shallow waters of sea grass. These waters are so shallow, and so clear, that a single diver may harvest many conch in a single day. Not only is conch meat a local delicacy and an important part of the local diet, but the large ornate shells are used in art and can be crafted into musical instruments. Because almost anyone with a small boat, snorkel, and mask, can participate in the conch harvest, it is essentially nonexcludable. At the same time, fishing for conch is rivalrous; once a diver catches one conch it cannot be caught by another diver.

Goods that are nonexcludable and rivalrous are called common resources. Because the waters of the Caribbean are open to all conch fishermen, and because any conch that you catch is conch that I cannot catch, common resources like the conch tend to be overharvested.

The problem of overharvesting common resources is not a new one, but ecologist Garret Hardin put the tag “Tragedy of the Commons” to the problem in a 1968 article in the magazine *Science*. Economists view this as a problem of property rights. Since nobody owns the ocean, or the conch that crawl on the sand beneath it, no one individual has an incentive to protect that resource and responsibly harvest it. To address the issue of overharvesting conch and other marine fisheries, economists typically advocate simple devices like fishing licenses, harvest limits, and shorter fishing seasons. When the population of a species drops to critically low numbers, governments have even banned the harvest until
Watch It

Watch the clip from this video to see more examples of common goods and the tragedy of the commons. Visit this page in your course online to view this presentation.

Try It

Visit this page in your course online to check your understanding.

Glossary

**non-excludable goods**: a good that no one is excluded from using

**non-rivalrous**: a good that does not get used up, meaning that if one person uses the public good, another can also use it

---

**FREE RIDERS**

Learning Objectives

- Explain the free rider problem

Free Riders

Private companies find it difficult to earn profits from producing public goods. If a good or service is nonexcludable, like national defense, so that it is impossible or very costly to exclude people from using this good or service, then how can a firm charge people for it? Once it's provided no one will voluntarily pay for it.

LINK IT UP

Visit this website to read about a connection between free riders and “bad music.”
When individuals make decisions about buying a public good, a free rider problem can arise, in which people have an incentive to let others pay for the public good and then to “free ride” on the purchases of others. The free rider problem can be expressed in terms of the prisoner’s dilemma game, which we learned about in the module on monopolistic competition and oligopoly. Say that two people are thinking about contributing to a public good: Rachel and Samuel. When either of them contributes to a public good, such as a local fire department, their personal cost of doing so is $4 and the social benefit of that person’s contribution is $6. Because society’s benefit of $6 is greater than the cost of $4, the investment is a good idea for society as a whole. The problem is that, while Rachel and Samuel pay for the entire cost of their contribution to the public good, they receive only half of the benefit, because the benefit of the public good is divided equally among the members of society. This sets up the prisoner’s dilemma illustrated in Table 1.

Table 1. Contributing to a Public Good as a Prisoner’s Dilemma

<table>
<thead>
<tr>
<th>Samuel (S) Contribute</th>
<th>Samuel (S) Do Not Contribute</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rachel (R) Contribute</td>
<td>R pays $4, receives $6, net gain +$2</td>
</tr>
<tr>
<td></td>
<td>S pays $4, receives $6, net gain +$2</td>
</tr>
<tr>
<td>Rachel (R) Do Not Contribute</td>
<td>R pays $0, receives $3, net gain +$3</td>
</tr>
<tr>
<td></td>
<td>S pays $4, receives $3, net gain −$1</td>
</tr>
</tbody>
</table>

If neither Rachel nor Samuel contributes to the public good, then there are no costs and no benefits of the public good. Suppose, however, that only Rachel contributes, while Samuel does not. Rachel incurs a cost of $4, but receives only $3 of benefit (half of the total $6 of benefit to society), while Samuel incurs no cost, and yet he also receives $3 of benefit. In this outcome, Rachel actually loses $1 while Samuel gains $3. A similar outcome, albeit with roles reversed, would occur if Samuel had contributed, but Rachel had not. Finally, if both parties contribute, then each incurs a cost of $4 and each receives $6 of benefit (half of the total $12 benefit to society). There is a dilemma with the Prisoner’s Dilemma, though, as you can see in the following example.

THE PROBLEM WITH THE PRISONER’S DILEMMA

The difficulty with the prisoner’s dilemma arises as each person thinks through his or her strategic choices.

**Step 1.** Rachel reasons in this way: If Samuel does not contribute, then I would be a fool to contribute. However, if Samuel does contribute, then I can come out ahead by not contributing.

**Step 2.** Either way, I should choose not to contribute, and instead hope that I can be a free rider who uses the public good paid for by Samuel.

**Step 3.** Samuel reasons the same way about Rachel.

**Step 4.** When both people reason in that way, the public good never gets built, and there is no movement to the option where everyone cooperates—which is actually best for all parties.

Try It

Visit this page in your course online to check your understanding.

The Role of Government in Paying for Public Goods

The key insight in paying for public goods is to find a way of assuring that everyone will make a contribution and to prevent free riders. For example, if people come together through the political process and agree to pay taxes and make group decisions about the quantity of public goods, they can defeat the free rider problem by requiring, through the law, that everyone contributes.

However, government spending and taxes are not the only way to provide public goods. In some cases, markets can produce public goods. For example, think about radio. It is nonexcludable, since once the radio signal is being
Free riders: those who want others to pay for the public good and then plan to use the good themselves; if many people act as free riders, the public good may never be provided.

Free ride: a song written by Dan Hartman and performed by The Edgar Winter Group. The single, engineered by Jim Reeves, was a top 20 U.S. hit in 1973, hitting number 14 on the Billboard Hot 100 Chart.
INTRODUCTION TO POSITIVE AND NEGATIVE EXTERNALITIES

What you’ll learn to do: define and give examples of positive and negative externalities

Figure 1. What does society gain from investing in the education of another student? While not always easy to measure, recent studies suggest the positive externalities to education typically include better health outcomes for the population, lower levels of crime, a cleaner environment and a more stable, democratic government.

In this section, you will explore in detail the spillover effects of positive and negative externalities. These include positive effects like improved technologies or negative effects like pollution.

POSITIVE EXTERNALITIES AND TECHNOLOGY

Learning Objectives
Identify and explain positive externalities, including new technology
Show how differences between private benefits and social benefits cause market failure

Individual demand reflects the marginal benefits a consumer receives from some product. When we sum all individual demands, we get the market demand. Market demand captures the **marginal private benefits** (MPB) of the product, since it measures the benefits received by the consumers who purchase the product. Figure 1 shows the market demand curve as MPB.

Some products provide both private and external benefits. **External benefits** are benefits received by someone who didn't purchase the product, but received some benefits as a spillover or side effect of the consumer’s purchasing the product. When external benefits exist, we describe the situation as a **positive externality**, where the marginal benefit to society is greater than the marginal benefits to the consumers who purchased the product. The **marginal social benefits** (MSB) are the marginal private benefits plus the external benefits. Figure 1 shows the social demand curve as MSB. Failure to consider those external benefits results in a market failure. In this section we examine some examples.

**Positive Externalities and Private Benefits**

Market competition can provide an incentive for discovering new technology because a firm can earn higher profits by finding a way to produce products more cheaply or to create products with characteristics consumers want. As Gregory Lee, CEO of Samsung said, “Relentless pursuit of new innovation is the key principle of our business and enables consumers to discover a world of possibilities with technology.” An innovative firm knows that it will usually have a temporary edge over its competitors and thus an ability to earn above-normal profits before competitors can catch up.

In certain cases, however, competition can discourage new technology, especially when other firms can quickly copy a new idea. Consider a pharmaceutical firm deciding to develop a new drug. On average, it can cost $800 million and take more than a decade to discover a new drug, perform the necessary safety tests, and bring the drug to market. If the research and development (R&D) effort fails—and every R&D project has some chance of failure—then the firm will suffer losses and could even be driven out of business. If the project succeeds, then the firm's competitors may figure out ways of adapting and copying the underlying idea, but without having to pay the costs themselves. As a result, the innovative company will bear the much higher costs of the R&D and will enjoy at best only a small, temporary advantage over the competition.

Many inventors over the years have discovered that their inventions brought them less profit than they might have reasonably expected.
• Eli Whitney (1765–1825) invented the cotton gin, but then southern cotton planters built their own seed-separating devices with a few minor changes in Whitney's design. When Whitney sued, he found that the courts in southern states would not uphold his patent rights.

• Thomas Edison (1847–1931) still holds the record for most patents granted to an individual. His first invention was an automatic vote counter, and despite the social benefits, he could not find a government that wanted to buy it.

• Gordon Gould (1920–2005) came up with the idea behind the laser in 1957. He put off applying for a patent and, by the time he did apply, other scientists had laser inventions of their own. A lengthy legal battle resulted, in which Gould spent $100,000 on lawyers, before he eventually received a patent for the laser in 1977. Compared to the enormous social benefits of the laser, Gould received relatively little financial reward.

A variety of studies by economists have found that the original inventor receives one-third to one-half of the total economic benefits from innovations, while other businesses and new product users receive the rest.

The Positive Externalities of New Technology

Will private firms in a market economy under invest in research and technology? If a firm builds a factory or buys a piece of equipment, the firm receives all the economic benefits that result from the investments. However, when a firm invests in new technology, the private benefits, or profits, that the firm receives are only a portion of the overall social benefits. The social benefits of an innovation take into account the value of all the positive externalities of the new idea or product, whether enjoyed by other companies or society as a whole, as well as the private benefits received by the firm that developed the new technology. Positive externalities are beneficial spillovers to a third party, or parties.

Consider the example of the Big Drug Company, which is planning its R&D budget for the next year. Economists and scientists working for Big Drug have compiled a list of potential research and development projects and estimated rates of return. (The rate of return is the estimated payoff from the project.) Figure 2 shows how the calculations work. The downward-sloping $D_{\text{private}}$ curve represents the firm’s demand for financial capital and reflects the company’s willingness to borrow to finance research and development projects at various interest rates. Suppose that this firm’s investment in research and development creates a spillover benefit to other firms and households. After all, new innovations often spark other creative endeavors that society also values. If we add the spillover benefits society enjoys to the firm’s private demand for financial capital, we can draw $D_{\text{social}}$ that lies above $D_{\text{private}}$.

If there was a way for the firm to fully monopolize those social benefits by somehow making them unavailable to the rest of us, the firm’s private demand curve would be the same as society’s demand curve. According to Figure 2 and Table 1, if the going rate of interest on borrowing is 8%, and the company can receive the private benefits of innovation only, then the company would finance $30 million. Society, at the same rate of 8%, would find it optimal to have $52 million of...
borrowing. Unless there is a way for the company to fully enjoy the total benefits, then it will borrow less than the socially optimal level of $52 million.

### Table 1. Return and Demand for Capital

<table>
<thead>
<tr>
<th>Rate of Return</th>
<th>$D_{\text{Private}}$ (in millions)</th>
<th>$D_{\text{Social}}$ (in millions)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2%</td>
<td>$72</td>
<td>$84</td>
</tr>
<tr>
<td>4%</td>
<td>$52</td>
<td>$72</td>
</tr>
<tr>
<td>6%</td>
<td>$38</td>
<td>$62</td>
</tr>
<tr>
<td>8%</td>
<td>$30</td>
<td>$52</td>
</tr>
<tr>
<td>10%</td>
<td>$26</td>
<td>$44</td>
</tr>
</tbody>
</table>

Big Drug's original demand for financial capital ($D_{\text{Private}}$) is based on the profits received by the firm. However, other pharmaceutical firms and health care companies may learn new lessons about how to treat certain medical conditions and are then able to create their own competing products. The social benefit of the drug takes into account the value of all the positive externalities of the drug. If Big Drug were able to gain this social return instead of other companies, its demand for financial capital would shift to the demand curve $D_{\text{Social}}$, and it would be willing to borrow and invest $52 million. However, if Big Drug is receiving only 50 cents of each dollar of social benefits, the firm will not spend as much on creating new products. The amount it would be willing to spend would fall somewhere in between $D_{\text{Private}}$ and $D_{\text{Social}}$.

### Try It

Visit this page in your course online to check your understanding.

### Why Invest in Human Capital?

The investment in anything, whether it is the construction of a new power plant or research in a new cancer treatment, usually requires a certain upfront cost with an uncertain future benefit. The investment in education, or human capital, is no different. Over the span of many years, a student and her family invest significant amounts of time and money into that student’s education. The idea is that higher levels of educational attainment will eventually serve to increase the student’s future productivity and subsequent ability to earn. Once the numbers are crunched, does this investment pay off for the student?

Almost universally, economists have found that the answer to this question is a clear “Yes.” For example, several studies of the return to education in the United States estimate that the rate of return to a college education is approximately 10-15%. Data in Table 2, from the U.S. Bureau of Labor Statistics’ *Usual Weekly Earnings of Wage and Salary Workers, Third Quarter 2016*, demonstrate that median weekly earnings are higher for workers who have completed more education. While these rates of return will beat equivalent investments in Treasury bonds or savings accounts, the estimated returns to education go primarily to the individual worker, so these returns are private rates of return to education.

### Table 2. Usual Weekly Earnings of Wage and Salary Workers, Fourth Quarter 2016 (Source: http://www.bls.gov/news.release/pdf/wkyeng.pdf)

<table>
<thead>
<tr>
<th></th>
<th>Less than a High School Degree</th>
<th>High School Degree, No College</th>
<th>Bachelor’s Degree</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median Weekly Earnings (full-time workers over the age of 25)</td>
<td>$519</td>
<td>$698</td>
<td>$1,270</td>
</tr>
</tbody>
</table>
What does society gain from investing in the education of another student? After all, if the government is spending taxpayer dollars to subsidize public education, society should expect some kind of return on that spending. Economists like George Psacharopoulos have found that, across a variety of nations, the social rate of return on schooling is also positive. After all, positive externalities exist from investment in education. While not always easy to measure, according to Walter McMahon, the positive externalities to education typically include better health outcomes for the population, lower levels of crime, a cleaner environment and a more stable, democratic government. For these reasons, many nations have chosen to use taxpayer dollars to subsidize primary, secondary, and higher education. Education clearly benefits the person who receives it, but a society where most people have a good level of education provides positive externalities for all.

Other Examples of Positive Externalities

Although technology may be the most prominent example of a positive externality, it is not the only one. For example, being vaccinated against disease is not only a protection for the individual, but it has the positive spillover of protecting others who may become infected. When a number of homes in a neighborhood are modernized, updated, and restored, not only does it increase the value of those homes, but the value of other properties in the neighborhood may increase as well.

Figure 3 shows the market for flu shots. The market demand curve $D_{\text{Market}}$ for flu shots reflects only the marginal private benefits (MPB) that the vaccinated individuals receive from the shots. Assuming that there are no spillover costs in the production of flu shots, the market supply curve is given by the marginal private cost (MPC) of producing the vaccinations.

The equilibrium quantity of flu shots produced in the market, where MPB is equal to MPC, is $Q_{\text{Market}}$ and the price of flu shots is $P_{\text{Market}}$. However, spillover benefits exist in this market because others, those who chose not to purchase a flu shot, receive a positive externality in a reduced chance of contracting the flu. When we add the spillover benefits to the marginal private benefit of flu shots, the marginal social benefit (MSB) of flu shots is given by $D_{\text{Social}}$. Because the MSB is greater than MPB, we see that the socially optimal level of flu shots is greater than the market quantity ($Q_{\text{Social}}$ exceeds $Q_{\text{Market}}$) and the corresponding price of flu shots, if the market were to produce $Q_{\text{Social}}$, would be at $P_{\text{Social}}$. Unfortunately, the marketplace does not recognize the positive externality and flu shots will go under produced and under consumed.

How can government try to move the market level of output closer to the socially desirable level of output? One policy would be to provide a subsidy, like a voucher, to any citizen who wishes to get vaccinated. This voucher would act as “income” that one could use purchase only a flu shot and, if the voucher were exactly equal to the per-unit spillover benefits, would increase market equilibrium to a quantity of $Q_{\text{Social}}$ and a price of $P_{\text{Social}}$ where MSB equals MSC. Suppliers of the flu shots would receive payment of $P_{\text{Social}}$ per vaccination, while consumers of flu shots would redeem the voucher and only pay a price of $P_{\text{Subsidy}}$. When the government uses a subsidy in this way, it produces the socially optimal quantity of vaccinations.
Figure 3. The Market for Flu Shots with Spillover Benefits (A Positive Externality). The market demand curve does not reflect the positive externality of flu vaccinations, so only $Q_{\text{Market}}$ will be exchanged. This outcome is inefficient because the marginal social benefit exceeds the marginal social cost. If the government provides a subsidy to consumers of flu shots, equal to the marginal social benefit minus the marginal private benefit, the level of vaccinations can increase to the socially optimal quantity of $Q_{\text{Social}}$.

Watch It

Watch this video to see how positive externalities increase the overall social benefits to society and cause the supply and demand graph to change.
Watch this video online: https://youtu.be/TSTLLFJbaA4

Try It

These questions allow you to get as much practice as you need, as you can click the link at the top of the first question (“Try another version of these questions”) to get a new set of questions. Practice until you feel comfortable doing the questions.
Visit this page in your course online to practice before taking the quiz.

Glossary

- **external benefits (or positive externalities)**: beneficial spillovers to a third party of parties, who did not purchase the good or service that provided the externalities

- **marginal private benefits (MPB)**: the benefits obtained by consumers from purchasing additional units of some product; shown by the market demand curve

- **marginal social benefits (MSP)**: the sum of the private and external benefits when additional units of some product are purchased; also known as the social demand curve

- **positive externalities**: beneficial spillovers to a third party or parties

- **private benefits**: the benefits a person who consumes a good or service receives, or a new product’s benefits or process that a company invents that the company captures

- **private rates of return**: 
when the estimated rates of return go primarily to an individual; for example, earning interest on a savings account

**social benefits**: the sum of private benefits and external benefits

**social rate of return**: when the estimated rates of return go primarily to society; for example, providing free education

---

**NEGATIVE EXTERNALITIES: POLLUTION**

**Learning Objectives**

- Explain and give examples of negative externalities, including pollution
- Show how differences between private costs and social costs cause market failure

A negative externality exists when the cost to society of an economic agent’s action is greater than the cost to the agent. In other words, there are external costs. Failure to consider those external costs results in a market failure. In this section we examine some examples.

**Externalities and Pollution**

In 1969, the Cuyahoga River in Ohio was so polluted that it spontaneously burst into flame. Air pollution was so bad at that time that Chattanooga, Tennessee was a city where, as an article from Sports Illustrated put it: “the death rate from tuberculosis was double that of the rest of Tennessee and triple that of the rest of the United States, a city in which the filth in the air was so bad it melted nylon stockings off women’s legs, in which executives kept supplies of clean white shirts in their offices so they could change when a shirt became too gray to be presentable, in which headlights were turned on at high noon because the sun was eclipsed by the gunk in the sky.”
The problem of pollution arises for every economy in the world, whether high-income or low-income, and whether market-oriented or command-oriented. Every country needs to strike some balance between production and environmental quality. This module begins by discussing how firms may fail to take certain social costs, like pollution, into their planning if they do not need to pay these costs. Traditionally, policies for environmental protection have focused on governmental limits on how much of each pollutant could be emitted. While this approach has had some success, economists have suggested a range of more flexible, market-oriented policies that reduce pollution at a lower cost. We will consider both approaches, but first let’s see how economists frame and analyze these issues.

**Keystone XL**

You might have heard about Keystone XL in the news. It is a pipeline system designed to bring oil from Canada to the refineries near the Gulf of Mexico, as well as to boost crude oil production in the United States. While a private company, TransCanada, will own the pipeline, U.S. government approval is required because of its size and location. The pipeline is being built in four phases, with the first two currently in operation, bringing oil from Alberta, Canada, east across Canada, south through the United States into Nebraska and Oklahoma, and northeast again to Illinois. The third and fourth phases of the project, known as Keystone XL, would create a pipeline southeast from Alberta straight to Nebraska, and then from Oklahoma to the Gulf of Mexico.

Sounds like a great idea, right? A pipeline that would move much needed crude oil to the Gulf refineries would increase oil production for manufacturing needs, reduce price pressure at the gas pump, and increase overall economic growth. Supporters argue that the pipeline is one of the safest pipelines built yet, and would reduce America’s dependence on politically vulnerable Middle Eastern oil imports. Not so fast, say its critics. The Keystone XL would be constructed over an enormous aquifer (one of the largest in the world) in the Midwest, and through an environmentally fragile area in Nebraska, causing great concern among environmentalists about possible destruction to the natural surroundings. They argue that leaks could taint valuable water sources and construction of the pipeline could disrupt and even harm indigenous species. Environmentalist groups have fought government approval of the proposed construction of the pipeline, and as of press time the pipeline projects remain stalled.

Of course, environmental concerns matter when discussing issues related to economic growth. But how much should they factor in? In the case of the pipeline, how do we know how much damage it would cause when we do not know how to put a value on the environment? Would the benefits of the pipeline outweigh the opportunity cost? The issue of how to balance economic progress with unintended effects on our planet is the subject of this module.

**Try It**

Visit this page in your course online to check your understanding.
The Economics of Pollution

From 1970 to 2012, the U.S. population increased by one-third and the size of the U.S. economy more than doubled. Since the first Earth Day in April 1970, the United States, using a variety of anti-pollution policies, has made genuine progress against a number of pollutants. Table 1 lists users of energy—from residential to industrial—the types of fuels each used, and the emissions from each, according to the U.S. Energy Information Administration (EIA). The table shows that emissions of certain key air pollutants declined substantially from 2007 to 2012; they dropped 730 million metric tons (MMT) a year—a 12% reduction. This seems to indicate that progress has been made in the United States in reducing overall carbon dioxide emissions, which cause greenhouse gases.

Table 1. Carbon Dioxide Emissions From Energy Consumption: (Million Metric Tons of Carbon Dioxide)

<table>
<thead>
<tr>
<th></th>
<th>Coal</th>
<th>Petroleum</th>
<th>Natural Gas</th>
<th>Purchased Electric Power</th>
<th>Total Energy Residential Sector CO2 Emissions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Residential Sector</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2007 December</td>
<td>0.082</td>
<td>11.237</td>
<td>38.986</td>
<td>78.282</td>
<td>128.587</td>
</tr>
<tr>
<td>2017 December</td>
<td>0</td>
<td>7.368</td>
<td>45.286</td>
<td>60.115</td>
<td>112.77</td>
</tr>
<tr>
<td>Change</td>
<td>-0.082</td>
<td>-3.869</td>
<td>6.3</td>
<td>-18.167</td>
<td>-15.817</td>
</tr>
<tr>
<td><strong>Commercial Sector</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2007 December</td>
<td>0.74</td>
<td>6.244</td>
<td>21.479</td>
<td>70.614</td>
<td>99.077</td>
</tr>
<tr>
<td>2017 December</td>
<td>0.169</td>
<td>6.543</td>
<td>26.818</td>
<td>53.671</td>
<td>87.201</td>
</tr>
<tr>
<td>Change</td>
<td>-0.571</td>
<td>0.299</td>
<td>5.339</td>
<td>-16.943</td>
<td>-11.876</td>
</tr>
<tr>
<td><strong>Industrial Sector</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2007 December</td>
<td>14.764</td>
<td>34.442</td>
<td>38.373</td>
<td>55.823</td>
<td>143.805</td>
</tr>
<tr>
<td>Change</td>
<td>-3.91</td>
<td>-6.333</td>
<td>10.022</td>
<td>-17.948</td>
<td>-18.965</td>
</tr>
<tr>
<td><strong>Transportation Sector</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2007 December</td>
<td>0</td>
<td>163.66</td>
<td>3.66</td>
<td>0.442</td>
<td>167.762</td>
</tr>
<tr>
<td>2017 December</td>
<td>0</td>
<td>155.442</td>
<td>4.485</td>
<td>0.327</td>
<td>160.254</td>
</tr>
<tr>
<td>Change</td>
<td>0</td>
<td>-8.218</td>
<td>0.825</td>
<td>-0.115</td>
<td>-7.508</td>
</tr>
<tr>
<td><strong>Power Sector</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Despite the gradual reduction in emissions from fossil fuels, many important environmental issues remain. Along with the still high levels of air and water pollution, other issues include hazardous waste disposal, destruction of wetlands and other wildlife habitats, and the impact on human health from pollution.

**Watch It**

Although this short clip was filmed over ten years ago, emissions remain a major problem and concern for China and the global economy.

Watch this video online: [https://youtu.be/vgfctghfDVo](https://youtu.be/vgfctghfDVo)

**Pollution as a Negative Externality**

Pollution is a negative externality. Economists illustrate the social costs of production with a demand and supply diagram. The **social costs** include the **private costs** of production incurred by the company and the external costs of pollution that are passed on to society. Figure 2 shows the demand and supply for manufacturing refrigerators. The demand curve (D) shows the quantity demanded at each price. The supply curve (S_{private}) shows the quantity of refrigerators supplied by all the firms at each price if they are taking only their private costs into account and they are allowed to emit pollution at zero cost. The market equilibrium (E_0), where quantity supplied and quantity demanded are equal, is at a price of $650 and a quantity of 45,000.
If the firm takes only its own costs of production into account, then its supply curve will be $S_{\text{private}}$, and the market equilibrium will occur at $E_0$. Accounting for additional external costs of $100 for every unit produced, the firm’s supply curve will be $S_{\text{social}}$. The new equilibrium will occur at $E_1$.

This information is also reflected in the first three columns of Table 2.

Table 2. A Supply Shift Caused by Pollution Costs

<table>
<thead>
<tr>
<th>Price</th>
<th>Quantity Demanded</th>
<th>Quantity Supplied before Considering Pollution Cost</th>
<th>Quantity Supplied after Considering Pollution Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>$600</td>
<td>50,000</td>
<td>40,000</td>
<td>30,000</td>
</tr>
<tr>
<td>$650</td>
<td>45,000</td>
<td>45,000</td>
<td>35,000</td>
</tr>
<tr>
<td>$700</td>
<td>40,000</td>
<td>50,000</td>
<td>40,000</td>
</tr>
<tr>
<td>$750</td>
<td>35,000</td>
<td>55,000</td>
<td>45,000</td>
</tr>
<tr>
<td>$800</td>
<td>30,000</td>
<td>60,000</td>
<td>50,000</td>
</tr>
<tr>
<td>$850</td>
<td>25,000</td>
<td>65,000</td>
<td>55,000</td>
</tr>
<tr>
<td>$900</td>
<td>20,000</td>
<td>70,000</td>
<td>60,000</td>
</tr>
</tbody>
</table>

However, as a by-product of the metals, plastics, chemicals and energy that are used in manufacturing refrigerators, some pollution is created. Let’s say that, if these pollutants were emitted into the air and water, they would create costs of $100 per refrigerator produced. These costs might occur because of injuries to human health, property values, wildlife habitat, reduction of recreation possibilities, or because of other negative impacts. In a market with no anti-pollution restrictions, firms can dispose of certain wastes absolutely free. Now imagine that firms which produce refrigerators must factor in these external costs of pollution—that is, the firms have to consider not only the costs of labor and materials needed to make a refrigerator, but also the broader costs to society of injuries to health and other values caused by pollution. If the firm is required to pay $100 for the additional external costs of pollution each time it produces a refrigerator, production becomes more costly and the entire supply curve shifts up by $100.

As illustrated in the fourth column of Table 2 and in Figure 2, the firm will need to receive a price of $700 per refrigerator and produce a quantity of 40,000—and the firm’s new supply curve will be $S_{\text{social}}$. The new equilibrium will occur at $E_1$, taking the additional external costs of pollution into account results in a higher price, a lower quantity of production, and
a lower quantity of pollution. The following feature will walk you through an example, this time with musical accompaniment.

**IDENTIFYING THE EQUILIBRIUM PRICE AND QUANTITY**

Table 3 shows the supply and demand conditions for a firm that will play trumpets on the streets when requested. Output is measured as the number of songs played.

<table>
<thead>
<tr>
<th>Price</th>
<th>Quantity Demanded</th>
<th>Quantity Supplied without paying the costs of the externality</th>
<th>Quantity Supplied after paying the costs of the externality</th>
</tr>
</thead>
<tbody>
<tr>
<td>$20</td>
<td>0</td>
<td>10</td>
<td>8</td>
</tr>
<tr>
<td>$18</td>
<td>1</td>
<td>9</td>
<td>7</td>
</tr>
<tr>
<td>$15</td>
<td>2.5</td>
<td>7.5</td>
<td>5.5</td>
</tr>
<tr>
<td>$12</td>
<td>4</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>$10</td>
<td>5</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>$5</td>
<td>7.5</td>
<td>2.5</td>
<td>0.5</td>
</tr>
</tbody>
</table>

**Step 1.** Determine the negative externality in this situation. To do this, you must think about the situation described and consider all parties that might be impacted. A negative externality might be the increase in noise pollution in the area where the firm is playing.

**Step 2.** Identify the equilibrium price and quantity when only private costs are taken into account, and then when social costs are taken into account. Remember that equilibrium is where the quantity demanded is equal to the quantity supplied.

**Step 3.** Look down the columns to where the quantity demanded (the second column) is equal to the “quantity supplied without paying the costs of the externality” (the third column). Then refer to the first column of that row to determine the equilibrium price. In this case, the equilibrium price and quantity when only private costs are taken into account would be at a price of $10 and a quantity of five.

**Step 4.** Identify the equilibrium price and quantity when the additional external costs are taken into account. Look down the columns of quantity demanded (the second column) and the “quantity supplied after paying the costs of the externality” (the fourth column) then refer to the first column of that row to determine the equilibrium price. In this case, the equilibrium will be at a price of $12 and a quantity of four.

**Step 5.** Consider how taking the externality into account affects the equilibrium price and quantity. Do this by comparing the two equilibrium situations. If the firm is forced to pay its additional external costs, then production of trumpet songs becomes more costly, and the supply curve will shift up.

**Try It**

Visit this page in your course online to check your understanding.

**Watch It**

Watch this video to review what happens to the supply and demand graph when a negative externality (in this case, plastic bags) results in greater social costs.

Watch this video online: [https://youtu.be/nBw6KvU51BE](https://youtu.be/nBw6KvU51BE)

Remember that the supply curve is based on choices about production that firms make while looking at their marginal costs, while the demand curve is based on the benefits that individuals perceive while maximizing utility. If no
externalities existed, private costs would be the same as the costs to society as a whole, and private benefits would be the same as the benefits to society as a whole. Thus, if no externalities existed, the interaction of demand and supply will equate marginal social costs and benefits, and thus produce the allocatively efficient outcome.

However, when the externality of pollution exists, the supply curve no longer represents all social costs. Because negative externalities represent a case where markets no longer consider all social costs, but only some of them, economists commonly refer to externalities as an example of market failure, as we learned about earlier. When there is market failure, the private market fails to achieve the efficient output, because either firms do not account for all costs incurred in the production of output and/or consumers do not account for all benefits obtained (a positive externality). In the case of pollution, at the market output, social costs of production exceed social benefits to consumers, and the market produces too much of the product.

We can see a general lesson here. If firms were required to pay the social costs of pollution, they would create less pollution but produce less of the product and charge a higher price.

Try It

These questions allow you to get as much practice as you need, as you can click the link at the top of the first question (“Try another version of these questions”) to get a new set of questions. Practice until you feel comfortable doing the questions.

Visit this page in your course online to practice before taking the quiz.

Glossary

private costs: production costs incurred by a company

social costs: the costs of production plus the external costs that are passed on to society

INTRODUCTION TO GOVERNMENT INVOLVEMENT WITH EXTERNALITIES

What you'll learn to do: analyze the efficacy of government policies to lessen positive and negative externalities
In the late 1960s and early 1970s the United States started passing comprehensive environmental laws that can be described as Command-and-Control Regulation. The question that persists is: “How effective has this regulation been?

With an understanding of what exactly positive and negative externalities are, we can now examine how the government uses regulation and economic incentives to help markets produce more efficient outcomes.

Learning Objectives

- Explain and give examples of command-and-control regulation

Command-and-Control Regulation

When the United States started passing comprehensive environmental laws in the late 1960s and early 1970s, a typical law specified how much pollution could be emitted out of a smokestack or a drainpipe and imposed penalties if that limit was exceeded. Other laws required the installation of certain equipment—for example, on automobile tailpipes or on smokestacks—to reduce pollution. These types of laws, which specify allowable quantities of pollution and which also may detail which pollution-control technologies must be used, fall under the category of effluent standards. In effect, effluent standards require that firms increase their costs by installing anti-pollution equipment; firms are thus required to take the social costs of pollution into account.
Command-and-control regulations, such as effluent standards, have been highly successful in protecting and cleaning up the U.S. environment. In 1970, the Environmental Protection Agency (EPA) was created to oversee all environmental laws. In the same year, the Clean Air Act was enacted to address air pollution. Just two years later, in 1972, Congress passed and the president signed the far-reaching Clean Water Act. These command-and-control environmental laws, and their amendments and updates, have been largely responsible for America’s cleaner air and water in recent decades. However, economists have pointed out three difficulties with command-and-control environmental regulation.

First, command-and-control regulation offers no incentive to improve the quality of the environment beyond the standard set by a particular law. Once the command-and-control regulation has been satisfied, polluters have zero incentive to do better.

Second, command-and-control regulation is inflexible. It usually requires the same standard for all polluters, and often the same pollution-control technology as well. This means that command-and-control regulation draws no distinctions between firms that would find it easy and inexpensive to meet the pollution standard—or to reduce pollution even further—and firms that might find it difficult and costly to meet the standard. Firms have no reason to rethink their production methods in fundamental ways that might reduce pollution even more and at lower cost.

Third, command-and-control regulations are written by legislators and the EPA, and so they are subject to compromises in the political process. Existing firms often argue (and lobby) that stricter environmental standards should not apply to them, only to new firms that wish to start production. Consequently, real-world environmental laws are full of fine print, loopholes, and exceptions.

Although critics accept the goal of reducing pollution, they question whether command-and-control regulation is the best way to design policy tools for accomplishing that goal.

---

**Watch It**

This video explains command and control regulations and takes a deeper look at when and why these types of policies are effective.
Visit this page in your course online to view this presentation.

**Try It**

Visit this page in your course online to check your understanding.

**Try It**

These questions allow you to get as much practice as you need, as you can click the link at the top of the first question (“Try another version of these questions”) to get a new set of questions. Practice until you feel comfortable doing the questions.
Visit this page in your course online to practice before taking the quiz.

**Glossary**

**effluent standard**: environmental regulation that limits pollution to a certain maximum quantity
Learning Objectives

- Evaluate the benefits and costs of environmental protection
- Apply marginal analysis to illustrate the marginal costs and marginal benefits of reducing pollution

Benefits and Costs of Clean Air and Clean Water

Government economists have estimated that U.S. firms may pay more than $200 billion per year to comply with federal environmental laws. That is a sizable amount of money. Is the money well spent?

The benefits of a cleaner environment can be divided into four areas: (1) people may stay healthier and live longer; (2) certain industries that rely on clean air and water, such as farming, fishing, and tourism, may benefit; (3) property values may be higher; and (4) people may simply enjoy a cleaner environment in a way that does not need to involve a market transaction. Some of these benefits, such as gains to tourism or farming, are relatively easy to value in economic terms. It is harder to assign a monetary value to others, such as the value of clean air for someone with asthma. It seems impossible to put a clear-cut monetary value on still others, such as the satisfaction you might feel from knowing that the air is clear over the Grand Canyon, even if you have never visited the Grand Canyon.

Although estimates of environmental benefits are not precise, they can still be revealing. For example, a study by the Environmental Protection Agency looked at the costs and benefits of the Clean Air Act from 1970 to 1990. It found that total costs over that time period were roughly $500 billion—a huge amount. However, it also found that a middle-range estimate of the health and other benefits from cleaner air was $22 trillion—about 44 times higher than the costs. A more recent study by the EPA estimated that the environmental benefits to Americans from the Clean Air Act will exceed their costs by a margin of four to one. The EPA estimated that “in 2010 the benefits of Clean Air Act programs will total about $110 billion. This estimate represents the value of avoiding increases in illness and premature death which would have prevailed.” Saying that overall benefits of environmental regulation have exceeded costs in the past, however, is very different from saying that every environmental regulation makes sense. For example, studies suggest that when breaking down emission reductions by type of contaminants, the benefits of air pollution control outweigh the costs primarily for particulates and lead, but when looking at other air pollutants, the costs of reducing them may be comparable to or greater than the benefits. Just because some environmental regulations have had benefits much higher than costs does not prove that every individual regulation is a sensible idea.

Try It

Visit this page in your course online to check your understanding.

Ecotourism: Making Environmentalism Pay
The definition of **ecotourism** is a little vague. Does it mean sleeping on the ground, eating roots, and getting close to wild animals? Does it mean flying in a helicopter to shoot anesthetic darts at African wildlife? Or a little of both? The definition may be fuzzy, but tourists who hope to appreciate the ecology of their destination—"eco tourists"—are the impetus to a big and growing business. The International Ecotourism Society estimates that international tourists interested in seeing nature or wildlife will take 1.56 billion trips by 2020.

**LINK IT UP**

Visit The International Ecotourism Society’s website to learn more about The International Ecotourism Society, its programs, and tourism’s role in sustainable community development.

Realizing the attraction of ecotourism, the residents of low-income countries may come to see that preserving wildlife habitats is more lucrative than, say, cutting down forests or grazing livestock to survive. In South Africa, Namibia, and Zimbabwe, for example, a substantial expansion of both rhinoceros and elephant populations is broadly credited to ecotourism, which has given local communities an economic interest in protecting them. Some of the leading ecotourism destinations include: Costa Rica and Panama in Central America; the Caribbean; Malaysia, and other South Pacific destinations; New Zealand; the Serengeti in Tanzania; the Amazon rain forests; and the Galapagos Islands. In many of these countries and regions, governments have enacted policies whereby revenues from ecotourism are shared with local communities, to give people in those local communities a kind of property right that encourages them to conserve their local environment.

Ecotourism needs careful management, so that the combination of eager tourists and local entrepreneurs does not destroy what the visitors are coming to see. But whatever one's qualms are about certain kinds of ecotourism—such as the occasional practice of rich tourists shooting elderly lions with high-powered rifles—it is worth remembering that the alternative is often that low-income people in poor countries will damage their local environment in their effort to survive.

**Marginal Benefits and Marginal Costs**

We can use the tools of marginal analysis to illustrate the marginal costs and the marginal benefits of reducing pollution. Figure 1 illustrates a theoretical model of this situation.

![Figure 1. Marginal Costs and Marginal Benefits of Environmental Protection. Reducing pollution is costly—resources must be sacrificed. The marginal costs of reducing pollution are generally increasing, because the least expensive and easiest reductions can be made first, leaving the more expensive methods for later. The marginal benefits of reducing pollution are generally declining, because the steps that provide the greatest benefit can be taken first, and steps that provide less benefit can wait until later.](image-url)
When the quantity of environmental protection is low so that pollution is extensive—for example, at quantity Qa—there are usually a lot of relatively cheap and easy ways to reduce pollution, and the marginal benefits of doing so are quite high. At Qa, it makes sense to allocate more resources to fight pollution. However, as the extent of environmental protection increases, the cheap and easy ways of reducing pollution begin to decrease, and more costly methods must be used. The marginal cost curve rises. Also, as environmental protection increases, the largest marginal benefits are achieved first, followed by reduced marginal benefits. As the quantity of environmental protection increases to, say, Qb, the gap between marginal benefits and marginal costs narrows. At point Qc the marginal costs will exceed the marginal benefits. At this level of environmental protection, society is not allocating resources efficiently, because too many resources are being given up to reduce pollution.

As society draws closer to Qb, some might argue that it becomes more important to use market-oriented environmental tools to hold down the costs of reducing pollution. Their objective would be to avoid environmental rules that would provide the quantity of environmental protection at Qc, where marginal costs exceed marginal benefits. The following feature delves into how the EPA measures its policies— and the monetary value of our lives.

Try It
Visit this page in your course online to check your understanding.

WHAT’S A LIFE WORTH?

The U.S. Environmental Protection Agency (EPA) must estimate the value of saving lives by reducing pollution against the additional costs. In measuring the benefits of government environmental policies, the EPA’s National Center for Environmental Economics (NCEE) values a statistical human life at $9.1 million (in 2011 U.S. dollars). Economists often value a human life on the basis of what income an individual could expect to earn over their lifetime. They may use studies of the value that people actually place on human lives in their own decisions. For example, some jobs have a higher probability of death than others, and these jobs typically pay more to compensate for the risk. Examples are ocean fishery as opposed to fish farming, and ice trucking in Alaska as opposed to truck driving in the “lower forty-eight” states.

Government regulators use estimates such as these when deciding what proposed regulations are “reasonable,” which means deciding which proposals have high enough benefits to justify their cost. For example, when the U.S. Department of Transportation makes decisions about what safety systems should be required in cars or airplanes, it will approve rules only where the estimated cost per life saved is $3 million or less.

Resources spent on life-saving regulations create tradeoff. A study by W. Kip Viscusi of Vanderbilt University estimated that when a regulation costs $50 million, it diverts enough spending in the rest of the economy from health care and safety expenditures that it costs a life. This finding suggests that any regulation that costs more than $50 million per life saved actually costs lives, rather than saving them.

Watch It
This video reviews some of these concepts and explains that nuances if government involvement in trying to reduce pollution.
Watch this video online: https://youtu.be/j6TnMGJF9sA
How Governments Can Encourage Innovation

Learning Objectives

- Identify and explain ways that some U.S. government policies encourage innovation

Intellectual Property Rights

A number of different government policies can increase the incentives to innovate, including: guaranteeing intellectual property rights, government assistance with the costs of research and development, and cooperative research ventures between universities and companies.

One way to increase new technology is to guarantee the innovator an exclusive right to that new product or process. Intellectual property rights include patents, which give the inventor the exclusive legal right to make, use, or sell the invention for a limited time, and copyright laws, which give the author an exclusive legal right over works of literature, music, film/video, and pictures. For example, if a pharmaceutical firm has a patent on a new drug, then no other firm can manufacture or sell that drug for twenty-one years, unless the firm with the patent grants permission. Without a patent, the pharmaceutical firm would have to face competition for any successful products, and could earn no more than a normal rate of profit. With a patent, a firm is able to earn monopoly profits on its product for a period of time—which offers an incentive for research and development. In general, how long can “a period of time” be? The following paragraph discusses patent and copyright protection timeframes for some works you might have heard of.

HOW LONG IS MICKEY MOUSE PROTECTED FROM BEING COPIED?

All patents and copyrights are scheduled to end someday. In 2003, copyright protection for Mickey Mouse was scheduled to run out. Once the copyright had expired, anyone would be able to copy Mickey Mouse cartoons or draw and sell new ones. In 1998, however, Congress passed the Sonny Bono Copyright Term Extension Act. For copyrights owned by companies or other entities, it increased or extended the copyright from 75 years to 95 years after publication. For copyrights owned by individuals, it increased or extended the copyright coverage from 50 years to 70 years after death. Along with protecting Mickey for another 20 years, the copyright extension affected about 400,000 books, movies, and songs.

Figure 1 illustrates how the total number of patent applications filed with the U.S. Patent and Trademark Office, as well as the total number of patents granted, surged in the mid-1990s with the invention of the Internet, and is still going strong today.
While patents provide an incentive to innovate by protecting the innovator, they are not perfect. For example:

- In countries that already have patents, economic studies show that inventors receive only one-third to one-half of the total economic value of their inventions.
- In a fast-moving high-technology industry like biotechnology or semiconductor design, patents may be almost irrelevant because technology is advancing so quickly.
- Not every new idea can be protected with a patent or a copyright—for example, a new way of organizing a factory or a new way of training employees.
- Patents may sometimes cover too much or be granted too easily. In the early 1970s, Xerox had received over 1,700 patents on various elements of the photocopy machine. Every time Xerox improved the photocopier, it received a patent on the improvement.
- The 21-year time period for a patent is somewhat arbitrary. Ideally, a patent should cover a long enough period of time for the inventor to earn a good return, but not so long that it allows the inventor to charge a monopoly price permanently.

Because patents are imperfect and do not apply well to all situations, alternative methods of improving the rate of return for inventors of new technology are desirable. Some of these possible alternative policies are described in the following sections.

Policy #1: Government Spending on Research and Development

If the private sector does not have sufficient incentive to carry out research and development, one possibility is for the government to fund such work directly. Government spending can provide direct financial support for research and development (R&D) done at colleges and universities, nonprofit research entities, and sometimes by private firms, as well as at government-run laboratories. While government spending on research and development produces technology that is broadly available for firms to use, it costs taxpayers money and can sometimes be directed more for political than for scientific or economic reasons.

The first column of Table 1 shows the sources of total U.S. spending on research and development; the second column shows the total dollars of R&D funding by each source. The third column shows that, relative to the total amount of funding, 26% comes from the federal government, about 67% of R&D is done by industry, and less than 3% is done by universities and colleges.

<table>
<thead>
<tr>
<th>Sources of R&amp;D Funding</th>
<th>Amount ($ billions)</th>
<th>Percent of the Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Federal government</td>
<td>$113.1</td>
<td>22.7%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sources of R&amp;D Funding</th>
<th>Amount ($ billions)</th>
<th>Percent of the Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Industry</td>
<td>$344.9</td>
<td>69.0%</td>
</tr>
<tr>
<td>Universities and colleges</td>
<td>$17.1</td>
<td>3.4%</td>
</tr>
<tr>
<td>Nonprofits</td>
<td>$19.9</td>
<td>4.0%</td>
</tr>
<tr>
<td>Nonfederal government</td>
<td>$4.0</td>
<td>0.8%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>$499</strong></td>
<td></td>
</tr>
</tbody>
</table>

**Policy #2: Tax Breaks for Research and Development**

In the 1960s the federal government paid for about two-thirds of the nation’s R&D. Over time, the U.S. economy has come to rely much more heavily on industry-funded R&D. The federal government has tried to focus its direct R&D spending on areas where private firms are not as active. One difficulty with direct government support of R&D is that it inevitably involves political decisions about which projects are worthy. The scientific question of whether research is worthwhile can easily become entangled with considerations like the location of the congressional district in which the research funding is being spent.

A complementary approach to supporting R&D that does not involve the government’s close scrutiny of specific projects is to give firms a reduction in taxes depending on how much research and development they do. The federal government refers to this policy as the research and experimentation (R&E) tax credit. According to the Treasury Department: “. . . the R&E Credit is also a cost-effective policy for stimulating additional private sector investment. Most recent studies find that each dollar of foregone tax revenue through the R&E Tax Credit causes firms to invest at least a dollar in R&D, with some studies finding a benefit to cost ratio of 2 or 2.96.”

**Policy #3 Cooperative Research**

State and federal governments support research in a variety of ways. For example, United for Medical Research, a coalition of groups that seek funding for the National Institutes of Health, (which is supported by federal grants), states: “NIH-supported research added $69 billion to our GDP and supported seven million jobs in 2011 alone.” Other institutions, such as the National Academy of Scientists and the National Academy of Engineers, receive federal grants for innovative projects. The Agriculture and Food Research Initiative (AFRI) at the United States Department of Agriculture awards federal grants to projects that apply the best science to the most important agricultural problems, from food safety to childhood obesity. Cooperation between government-funded universities, academies, and the private sector can spur product innovation and create whole new industries.

**Try It**

Visit this page in your course online to check your understanding.

**Glossary**

**intellectual property**: the body of law including patents, trademarks, copyrights, and trade secret law that protect the right of inventors to produce and sell their inventions
MARKET-ORIENTED ENVIRONMENTAL TOOLS

Learning Objectives

- Show how pollution charges or marketable permits can impact firm decisions
- Explain the significance of marketable permits and property rights (the Coase theorem)
- Evaluate which government and market policies are most appropriate for various situations

Market-oriented environmental policies create incentives for firms to reduce pollution. The three main categories of market-oriented approaches to pollution control are pollution charges, marketable permits, and better-defined property rights. All of these policy tools, discussed below, address the shortcomings of command-and-control regulation—albeit in different ways.

Effluent Charges

An effluent (or pollution) charge is a tax imposed on the quantity of pollution that a firm emits. Under an effluent charge system, firms are allowed to pollute, as long as they pay the charge for every unit of pollution. A pollution charge gives a profit-maximizing firm an incentive to figure out ways to reduce its emissions—as long as the marginal cost of reducing the emissions is less than the tax.

For example, consider a small firm that emits 50 pounds per year of small particles, such as soot, into the air. Particulate matter, as it is called, causes respiratory illnesses and also imposes costs on firms and individuals.

Figure 1 illustrates the marginal costs that a firm faces in reducing pollution. The marginal cost of pollution reduction, like most most marginal cost curves increases with output, at least in the short run. Reducing the first 10 pounds of particulate emissions costs the firm $300. Reducing the second 10 pounds would cost $500; reducing the third ten pounds would cost $900; reducing the fourth 10 pounds would cost $1,500; and the fifth 10 pounds would cost $2,500. This pattern for the costs of reducing pollution is common, because the firm can use the cheapest and easiest method to make initial reductions in pollution, but additional reductions in pollution become more expensive.
Imagine the firm now faces a pollution tax of $1,000 for every 10 pounds of particulates emitted. The firm has the choice of either polluting and paying the tax, or reducing the amount of particulates they emit and paying the cost of abatement as shown in the figure. How much will the firm pollute and how much will the firm abate? The first 10 pounds would cost the firm $300 to abate. This is substantially less than the $1,000 tax, so they will choose to abate. The second 10 pounds would cost $500 to abate, which is still less than the tax, so they will choose to abate. The third 10 pounds would cost $900 to abate, which is slightly less than the $1,000 tax. The fourth 10 pounds would cost $1,500, which is much more costly than paying the tax. As a result, the firm will decide to reduce pollutants by 30 pounds, because the marginal cost of reducing pollution by this amount is less than the pollution tax. With a tax of $1,000, the firm has no incentive to reduce pollution more than 30 pounds.

A firm that has to pay a pollution tax will have an incentive to figure out the least expensive technologies for reducing pollution. Firms that can reduce pollution cheaply and easily will do so to minimize their pollution taxes, whereas firms that will incur high costs for reducing pollution will end up paying the pollution tax instead. If the pollution tax applies to every source of pollution, then no special favoritism or loopholes are created for politically well-connected producers.

For an example of a pollution charge at the household level, consider two ways of charging for garbage collection. One method is to have a flat fee per household, no matter how much garbage a household produces. An alternative approach is to have several levels of fees, depending on how much garbage the household produces—and to offer lower or free charges for recyclable materials. As of 2006 (latest statistics available), the EPA had recorded over 7,000 communities that have implemented “pay as you throw” programs. When people have a financial incentive to put out less garbage and to increase recycling, they find ways of doing so.

A number of environmental policies are really pollution charges, although they often do not travel under that name. For example, the federal government and many state governments impose taxes on gasoline. We can view this tax as a charge on the air pollution that cars generate as well as a source of funding for maintaining roads. Indeed, gasoline taxes are far higher in most other countries than in the United States.

Similarly, the refundable charge of five or 10 cents that only 10 states have for returning recyclable cans and bottles works like a pollution tax that provides an incentive to avoid littering or throwing bottles in the trash. Compared with command-and-control regulation, a pollution tax reduces pollution in a more flexible and cost-effective way.

** Marketable Permits**

When a city or state government sets up a marketable permit program, it must start by determining the overall quantity of pollution it will allow as it tries to meet national pollution standards. Then, a number of permits allowing only this quantity of pollution are divided among the firms that emit that pollutant. These permits to pollute can be sold or given to firms free.
Now, add two more conditions. Imagine that these permits are designed to reduce total emissions over time. For example, a permit may allow emission of 10 units of pollution one year, but only nine units the next year, then eight units the year after that, and so on down to some lower level. In addition, imagine that these are marketable permits, meaning that firms can buy and sell them.

To see how marketable permits can work to reduce pollution, consider the four firms listed in Table 1. The table shows current emissions of lead from each firm. At the start of the marketable permit program, each firm receives permits to allow this level of pollution. However, these permits are shrinkable, and next year the permits allow the firms to emit only half as much pollution. Let’s say that in a year, Firm Gamma finds it easy and cheap to reduce emissions from 600 tons of lead to 200 tons, which means that it has permits that it is not using that allow emitting 100 tons of lead. Firm Beta reduces its lead pollution from 400 tons to 200 tons, so it does not need to buy any permits, and it does not have any extra permits to sell. However, although Firm Alpha can easily reduce pollution from 200 tons to 150 tons, it finds that it is cheaper to purchase permits from Gamma rather than to reduce its own emissions to 100. Meanwhile, Firm Delta did not even exist in the first period, so the only way it can start production is to purchase permits to emit 50 tons of lead.

The total quantity of pollution will decline. But the buying and selling of the marketable permits will determine exactly which firms reduce pollution and by how much. With a system of marketable permits, the firms that find it least expensive to do so will reduce pollution the most.

Table 1. How Marketable Permits Work

<table>
<thead>
<tr>
<th></th>
<th>Firm Alpha</th>
<th>Firm Beta</th>
<th>Firm Gamma</th>
<th>Firm Delta</th>
</tr>
</thead>
<tbody>
<tr>
<td>Current emissions—permits distributed free for this amount</td>
<td>200 tons</td>
<td>400 tons</td>
<td>600 tons</td>
<td>0 tons</td>
</tr>
<tr>
<td>How much pollution will these permits allow in one year?</td>
<td>100 tons</td>
<td>200 tons</td>
<td>300 tons</td>
<td>0 tons</td>
</tr>
<tr>
<td>Actual emissions one year in the future</td>
<td>150 tons</td>
<td>200 tons</td>
<td>200 tons</td>
<td>50 tons</td>
</tr>
<tr>
<td>Buyer or seller of marketable permit?</td>
<td>Buys permits for 50 tons</td>
<td>Doesn’t buy or sell permits</td>
<td>Sells permits for 100 tons</td>
<td>Buys permits for 50 tons</td>
</tr>
</tbody>
</table>

Another application of marketable permits occurred when the Clean Air Act was amended in 1990. The revised law sought to reduce sulfur dioxide emissions from electric power plants to half of the 1980 levels out of concern that sulfur dioxide was causing acid rain, which harms forests as well as buildings. In this case, the marketable permits the federal government issued were free of charge (no pun intended) to electricity-generating plants across the country, especially those that were burning coal (which produces sulfur dioxide). These permits were of the “shrinkable” type; that is, the amount of pollution allowed by a given permit declined with time.

Try It

Visit this page in your course online to check your understanding.

Watch It

Watch this video to learn more about how marketable permits offer an efficient way to reduce pollution.
Watch this video online: https://youtu.be/9tUb3MDrgEc

Better-Defined Property Rights

A clarified and strengthened idea of property rights can also strike a balance between economic activity and pollution. Ronald Coase (1910–2013), who won the 1991 Nobel Prize in economics, offered a vivid illustration of an externality: a railroad track running beside a farmer’s field where the railroad locomotive sometimes emits sparks and sets the field ablaze. Coase asked whose responsibility it was to address this spillover. Should the farmer be required to build a tall
fence alongside the field to block the sparks, or should the railroad be required to place a gadget on the locomotive’s smokestack to reduce the number of sparks?

Coase pointed out that one cannot resolve this issue until one clearly defines property rights—that is, the legal rights of ownership on which others are not allowed to infringe without paying compensation. Does the farmer have a property right not to have a field burned? Does the railroad have a property right to run its own trains on its own tracks? If neither party has a property right, then the two sides may squabble endlessly, doing nothing, and sparks will continue to set the field aflame. However, if either the farmer or the railroad has a well-defined legal responsibility, then that party will seek out and pay for the least costly method of reducing the risk that sparks will hit the field. The property right determines whether the farmer or the railroad pays the bills.

The property rights approach is highly relevant in cases involving endangered species. The U.S. government’s endangered species list includes about 1,000 plants and animals, and about 90% of these species live on privately owned land. The protection of these endangered species requires careful thinking about incentives and property rights. The discovery of an endangered species on private land has often triggered an automatic reaction from the government to prohibit the landowner from using that land for any purpose that might disturb the imperiled creatures. Consider the incentives of that policy: if you admit to the government that you have an endangered species, the government effectively prohibits you from using your land. As a result, rumors abounded of landowners who followed a policy of “shoot, shovel, and shut up” when they found an endangered animal on their land. Other landowners have deliberately cut trees or managed land in a way that they knew would discourage endangered animals from locating there.

HOW EFFECTIVE ARE MARKET-ORIENTED ENVIRONMENTAL POLICY TOOLS?

Environmentalists sometimes fear that market-oriented environmental tools are an excuse to weaken or eliminate strict limits on pollution emissions and instead to allow more pollution. It is true that if pollution charges are set very low or if marketable permits do not reduce pollution by very much then market-oriented tools will not work well. However, command-and-control environmental laws can also be full of loopholes or have exemptions that do not reduce pollution by much, either. The advantage of market-oriented environmental tools is not that they reduce pollution by more or less, but because of their incentives and flexibility, they can achieve any desired reduction in pollution at a lower cost to society.

A more productive policy would consider how to provide private landowners with an incentive to protect the endangered species that they find and to provide a habitat for additional endangered species. For example, the government might pay landowners who provide and maintain suitable habitats for endangered species or who restrict the use of their land to protect an endangered species. Again, an environmental law built on incentives and flexibility offers greater promise than a command-and-control approach when trying to oversee millions of acres of privately owned land.

Try It

Visit this page in your course online to check your understanding.

Watch It

Robert Coase’s analysis of property rights and externalities led to the Coase theorem, which argues that the market can effectively solve the problem with externalities as long as the transaction costs are low and property rights are clearly defined. Watch this video for an example of this as applied to bees and the important role they play in creating honey and pollinating crops. While it seems likely that there could be a shortage of bees for the purpose of pollinating, beekeepers and farmers have worked out a solution themselves using the market.

Watch this video online: https://youtu.be/00HPak2RLIQ

Applying Market-Oriented Environmental Tools

Market-oriented environmental policies are a tool kit. Specific policy tools will work better in some situations than in others. For example, marketable permits work best when a few dozen or a few hundred parties are highly interested in
trading, as in the cases of oil refineries that trade lead permits or electrical utilities that trade sulfur dioxide permits. However, for cases in which millions of users emit small amounts of pollution—such as emissions from car engines or unrecycled soda cans—and have no strong interest in trading, pollution charges will typically offer a better choice. We can also combine market-oriented environmental tools. We can view marketable permits as a form of improved property rights. Alternatively, the government could combine marketable permits with a pollution tax on any emissions not covered by a permit.

Alternatively, if the externality is such that even a small amount poses catastrophic danger, like the earlier smallpox example or nuclear waste concerns, a government-imposed effluent standard is likely preferable.

Try It

Visit this page in your course online to check your understanding.

Glossary

**effluent (or pollution) charge**: a tax imposed on the quantity of pollution that a firm emits; also called a pollution tax

**marketable permit program**: a permit that allows a firm to emit a certain amount of pollution; firms with more permits than pollution can sell the remaining permits to other firms

**property rights**: the legal rights of ownership on which others are not allowed to infringe without paying compensation

LEARN BY DOING: MARKET-ORIENTED ENVIRONMENTAL TOOLS

Try It

These questions allow you to get as much practice as you need, as you can click the link at the top of the first question (“Try another version of these questions”) to get a new set of questions. Practice until you feel comfortable doing the questions.

Visit this page in your course online to practice before taking the quiz.
In this module, we learned that governments can either promote the supply of positive externality goods or inhibit the supply of negative externality goods through regulation or through financial incentives—subsidies on the former and charges on the latter.

Let's return to one of the questions we posed at the beginning of this module. Why does it make sense for society to subsidize healthcare? It's because healthcare has external benefits—people at large benefit from living in a community of healthy people. Healthy people are less likely to spread disease to others. Prevention of disease often costs less than treatment. If someone in a community gets sick due to lack of access to regular healthcare, the rest of the community may end up paying more through taxes when the ill person shows up at the local emergency room or hospital.

Licensing & Attributions

CC licensed content, Original

Public domain content
MODULE 12: LABOR MARKETS

WHY IT MATTERS: LABOR MARKETS

Why understand labor markets?

Figure 1. What determines incomes? In the U.S., income is based on one’s value to an employer, which depends in part on education. (Credit: modification of work by AFL-CIO America’s Unions/Flickr Creative Commons and COD Newsroom/Flickr Creative Commons)

The Increasing Value of a College Degree

Working your way through college used to be fairly common in the United States. According to a 2015 study by the Georgetown Center on Education and the Workforce, 40% of college students work 30 hours or more per week.

At the same time, the cost of college seems to rise every year. The data show that the cost of tuition, fees, room and board has more than doubled since 1984. Thus, even full time employment may not be enough to cover college expenses anymore. Working full time at minimum wage—40 hours per week, 52 weeks per year—earns $15,080 before taxes, which is less than the $19,548 the College Board estimates it cost in 2016 for a year of college at a public university. The result of these costs is that student loan debt topped $1.3 trillion this year.

Despite these disheartening figures, the value of a bachelor’s degree has never been higher. How do we explain this? This module will tell us. We will learn about:

- The theory of labor markets
- How wages are determined in an imperfectly competitive labor market
- How unions affect wages and employment
- How labor market outcomes are determined under bilateral monopoly
- Theories of employment discrimination
How Immigration affects labor market outcomes

In a market economy like the United States, income comes from ownership of the means of production: resources or assets. More precisely, one’s income is a function of two things: the quantity of each resource one owns, and the value society places on those resources.

Recall from the module on production and costs that each factor of production has an associated factor payment. For the majority of us, the most important resource we own is our labor. Thus, most of our income is wages, salaries, commissions, tips and other types of labor income. Your labor income depends on how many hours you have to work and the wage rate an employer will pay you for those hours. At the same time, some people own real estate, which they can either use themselves or rent out to other users. Some people have financial assets like bank accounts, stocks and bonds, for which they earn interest, dividends or some other form of income. Each of these factor payments, like wages for labor and interest for financial capital, is determined in their respective factor markets.

INTRODUCTION TO LABOR MARKETS

What you’ll learn to do: analyze labor markets and how supply and demand interact to determine the market wage rate

Just as a firm determines the amount of a product or service to provide based on marginal costs, a firm also makes decisions about employing people based on the marginal productivity of their labor. In this section, you’ll learn about how firms determine how many people to hire.
LABOR MARKETS

Learning Objectives

- Describe labor markets; explain why the value of the marginal product of labor is the demand for labor

What is the labor market?

The labor market is the term that economists use for all the different markets for labor. There is no single labor market. Rather, there is a different market for every different type of labor. Labor differs by type of work (e.g. retail sales vs. scientist), skill level (entry level or more experienced), and geographic location (the market for administrative assistants is probably more local or regional than the market for university presidents). While each labor market is different, they all tend to respond to similar disturbances in similar ways. For example, when wages go up for one type of job in an industry, they tend to go up in other types of jobs too. When economists talk about the labor market, they are describing these similarities.

The labor market, like all markets, has a demand and a supply. Why do firms demand labor? Why is an employer willing to pay you for your work? It’s not because the employer likes you or is socially conscious. Rather, it’s because your labor is worth something to the employer—your work brings in revenues to the firm. How much is an employer willing to pay? That depends on the skills and experience you bring to the firm.

If a firm wants to maximize profits, it will never pay more (in terms of wages and benefits) for a worker than the value of his or her marginal productivity to the firm. We call this the first rule of labor markets.

Suppose a worker can produce two widgets per hour and the firm can sell each widget for $4 each. Then the worker is generating $8 per hour in revenues to the firm, and a profit-maximizing employer will pay the worker up to, but no more than, $8 per hour, because that is what the worker is worth to the firm.

Recall the definition of marginal product. Marginal product is the additional output a firm can produce by adding one more worker to the production process. Since employers often hire labor by the hour, we’ll define marginal product as the additional output the firm produces by adding one more worker hour to the production process. In this module, we assume that workers are homogeneous—they have the same background, experience and skills and they put in the same amount of effort. Thus, marginal product depends on the capital and technology with which workers have to work.

A typist can type more pages per hour with an electric typewriter than a manual typewriter, and he or she can type even more pages per hour with a personal computer and word processing software. A ditch digger can dig more cubic feet of dirt in an hour with a backhoe than with a shovel.

We can define the demand for labor as the marginal product of labor times the value of that output to the firm.

<table>
<thead>
<tr>
<th># Workers (L)</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 1. Marginal Product of Labor</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
On what does the value of each worker’s marginal product depend? If we assume that the employer sells its output in a perfectly competitive market, the value of each worker’s output will be the market price of the product. Thus,

\[ \text{Demand for Labor} = MP_L \times P = \text{Value of the Marginal Product of Labor} \]

We show this in Table 2, which is an expanded version of Table 1.

<table>
<thead>
<tr>
<th># Workers (L)</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>MP_L</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Price of Output</td>
<td>$4</td>
<td>$4</td>
<td>$4</td>
<td>$4</td>
</tr>
<tr>
<td>VMP_L</td>
<td>$16</td>
<td>$12</td>
<td>$8</td>
<td>$4</td>
</tr>
</tbody>
</table>

Note that the value of each additional worker is less than the ones who came before.

Thus, the demand for labor (that is, the value of the marginal product of labor is downward sloping as the firm hires additional labor.
Watch It

This video takes us through the example of a restaurant interested in hiring janitors. With clean facilities, a restaurant will make more money, but they must consider the cost of a janitor versus the benefit from their labor. Watch the selected clip from this video to see how this correlates to a supply and demand graph. Visit this page in your course online to view this presentation.

Try It

Visit this page in your course online to view this presentation.

Try It

These questions allow you to get as much practice as you need, as you can click the link at the top of the first question (“Try another version of these questions”) to get a new set of questions. Practice until you feel comfortable doing the questions. Visit this page in your course online to practice before taking the quiz.

Glossary

first rule of labor markets: an employer will never pay a worker more than the value of the worker’s marginal productivity to the firm

value of the marginal product of labor: the marginal product of an additional worker multiplied by the price of the firm’s output

Figure 2. Value of the Marginal Product of Labor. For firms operating in a competitive output market, the value of additional output sold is the price the firms receive for the output. Since $MP_L$ declines with additional labor employed, while that marginal product is worth the market price, the value of the marginal product declines as employment increases.
THE DEMAND FOR LABOR

Learning Objectives

- Explain and graph the demand for labor in perfectly competitive output markets
- Explain and graph the demand for labor in imperfectly competitive output markets
- Demonstrate how supply and demand interact to determine the market wage rate

Demand for Labor in Perfectly Competitive Output Markets

The question for any firm is how much labor to hire.

We can define a perfectly competitive labor market as one where firms can hire all the labor they wish at the going market wage. Think about secretaries in a large city. Employers who need secretaries can probably hire as many as they need if they pay the going wage rate.

Graphically, this means that firms face a horizontal supply curve for labor, as Figure 1 shows.

Given the market wage, profit maximizing firms hire workers up to the point where: \( W_{\text{mkt}} = VMP_L \)
Figure 1. Equilibrium Employment for Firms in a Competitive Labor Market. In a perfectly competitive labor market, firms can hire all the labor they want at the going market wage. Therefore, they hire workers up to the point $L_1$ where the going market wage equals the value of the marginal product of labor.

**Derived Demand**

Economists describe the demand for inputs like labor as a derived demand. Since the demand for labor is $\text{MPL} \times P$, it is dependent on the demand for the product the firm is producing. We show this by the $P$ term in the demand for labor. An increase in demand for the firm’s product drives up the product’s price, which increases the firm’s demand for labor. Thus, we derive the demand for labor from the demand for the firm’s output.

**Try It**

Visit this page in your course online to check your understanding.

**Demand for Labor in Imperfectly Competitive Output Markets**

If an employer does not sell its output in a perfectly competitive industry, it faces a downward sloping demand curve for output. This means that in order to sell additional output the firm must lower its price. This is true if the firm is a monopoly, but it’s also true if the firm is an oligopoly or monopolistically competitive. In this situation, the value of an additional unit of output sold is the marginal revenue, rather than the price. This means that a worker’s marginal product is valued by the marginal revenue, not the price. Thus, the demand for labor is the marginal product times the marginal revenue, which we call the **marginal revenue product**.

The Demand for Labor = $\text{MP}_L \times MR = \text{Marginal Revenue Product}$

<table>
<thead>
<tr>
<th># Workers (L)</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{MP}_L$</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Marginal Revenue</td>
<td>$4</td>
<td>$3</td>
<td>$2</td>
<td>$1</td>
</tr>
<tr>
<td>$\text{MRP}_L$</td>
<td>$16</td>
<td>$9</td>
<td>$4</td>
<td>$1</td>
</tr>
</tbody>
</table>
Figure 2. Marginal Revenue Product. For firms with some market power in their output market, the value of additional output sold is the firm’s marginal revenue. Since MPL declines with additional labor employed and since MR declines with additional output sold, the firm’s marginal revenue declines as employment increases.

Everything else remains the same as we described above in the discussion of the labor demand in perfectly competitive labor markets. Given the market wage, profit-maximizing firms will hire workers up to the point where the market wage equals the marginal revenue product, as Figure 3 shows.

Figure 3. Equilibrium Level of Employment for Firms with Market Power. For firms with market power in their output market, they choose the number of workers, \( L_2 \), where the going market wage equals the firm’s marginal revenue product. Note that since marginal revenue is less than price, the demand for labor for a firm which has market power in its output market is less than the demand for labor (\( L_1 \)) for a perfectly competitive firm. As a result, employment will be lower in an imperfectly competitive industry than in a perfectly competitive industry.

Do Profit-Maximizing Employers Exploit Labor?
If you look back at Figure 3, you will see that the firm only pays the last worker it hires what they’re worth to the firm. Every other worker brings in more revenue than the firm pays him or her. This has sometimes led to the claim that employers exploit workers because they do not pay workers what they are worth. Let’s think about this claim. The first worker is worth $x to the firm, and the second worker is worth $y, but why are they worth that much? It is because of the capital and technology with which they work. The difference between workers’ worth and their compensation goes to pay for the capital, and other inputs in the production process. The difference also goes to the employer’s profit, without which the firm would close and workers wouldn’t have a job. The firm may be earning excessive profits, but that is a different topic of discussion.

Try It
Visit this page in your course online to check your understanding.

What Determines the Going Market Wage Rate?

We learned earlier that the labor market has demand and supply curves like other markets. The demand for labor curve is a downward sloping function of the wage rate. The market demand for labor is the horizontal sum of all firms’ demands for labor. The supply for labor curve is an upward sloping function of the wage rate. This is because if wages for a particular type of labor increase in a particular labor market, people with appropriate skills may change jobs, and vacancies will attract people from outside the geographic area. The market supply for labor is the horizontal summation of all individuals’ supplies of labor.

![Figure 4. The Market Wage Rate. In a competitive labor market, the equilibrium wage and employment level are determined where the market demand for labor equals the market supply of labor.](image)

Like all equilibrium prices, the market wage rate is determined through the interaction of supply and demand in the labor market. Thus, we can see in Figure 4, the wage rate and number of workers hired in a competitive labor market.

Watch It
Watch this video for a nice overview of the labor market, and the ways that supply and demand interact to determine wages. The video will also introduce some of the key concepts we’ll discuss soon, including monopsonies, unions, discrimination, and minimum wage laws.

Watch this video online: [https://youtu.be/mWwXmH-n5Bo](https://youtu.be/mWwXmH-n5Bo)
Glossary

**collective bargaining**: negotiations between unions and a firm or firms

**labor union**: an organization of workers that negotiates with employers over wages and working conditions

**perfectly competitive labor market**: a labor market where neither suppliers of labor nor demanders of labor have any market power; thus, an employer can hire all the workers they would like at the going market wage

**marginal revenue product of labor**: the marginal product of an additional worker multiplied by the marginal revenue to the firm of the additional worker’s output

LEARN BY DOING: THE DEMAND FOR LABOR

Try It

These questions allow you to get as much practice as you need, as you can click the link at the top of the first question (“Try another version of these questions”) to get a new set of questions. Practice until you feel comfortable doing the questions.

Visit this page in your course online to practice before taking the quiz.

INTRODUCTION TO WAGES AND EMPLOYMENT IN LABOR MARKETS

What you’ll learn to do: explain how wages are determined when employers or employees hold labor market power
In the labor market, employers often hold most of the power in determining wages. This could especially happen if the firm has little or no competition in hiring employees, so workers have little alternative to accepting the wages that are offered. This is called a monopsony, and results in a lower level of employment and a lower equilibrium wage than what would be preferred by the competitive labor market.

Sometimes, however, the employees hold more of the power in determining wages. This is the case with labor unions, who negotiate wages through collective bargaining.

---

**LABOR MARKET POWER BY EMPLOYERS**

**Learning Objectives**

- Explain how wages and employment are determined in labor markets where employers have market power (monopsonies)

In the modules on market structure (i.e. perfectly competition, monopoly, oligopoly, and monopolistic competition), we observed that while economists use the theory of perfect competition as an ideal case of market structure, there are very few examples of perfectly competitive output markets in the real world. What about labor markets? How many labor markets are perfectly competitive? There are probably more examples of perfectly competitive labor markets than perfectly competitive product markets, but that doesn’t mean that all labor markets are competitive.

When a job applicant is bargaining with an employer for a position, the applicant is often at a disadvantage—needing the job more than the employer needs that particular applicant. John Bates Clark (1847–1938), often named as the first great American economist, wrote in 1907: “In the making of the wages contract the individual laborer is always at a disadvantage. He has something which he is obliged to sell and which his employer is not obliged to take, since he [that is, the employer] can reject single men with impunity.”

To give workers more power, the U.S. government has passed, in response to years of labor protests, a number of laws to create a more equal balance of power between workers and employers. These laws include some of the following:

- Setting minimum hourly wages
- Setting maximum hours of work (at least before employers pay overtime rates)
- Prohibiting child labor
- Regulating health and safety conditions in the workplace
- Preventing discrimination on the basis of race, ethnicity, gender, sexual orientation, and age
- Requiring employers to provide family leave
- Requiring employers to give advance notice of layoffs
- Covering workers with unemployment insurance
- Setting a limit on the number of immigrant workers from other countries

Table 1 lists some prominent U.S. workplace protection laws. Many of the laws listed in the table were only the start of labor market regulations in these areas and have been followed, over time, by other related laws, regulations, and court rulings.

<table>
<thead>
<tr>
<th>Law</th>
<th>Protection</th>
</tr>
</thead>
<tbody>
<tr>
<td>National Labor-Management Relations Act of 1935 (the “Wagner Act”)</td>
<td>Establishes procedures for establishing a union that firms are obligated to follow; sets up the National Labor Relations Board for deciding disputes</td>
</tr>
<tr>
<td>Social Security Act of 1935</td>
<td>Under Title III, establishes a state-run system of unemployment insurance, in which workers pay into a state fund when they are employed and received benefits for a time when they are unemployed</td>
</tr>
<tr>
<td>Fair Labor Standards Act of 1938</td>
<td>Establishes the minimum wage, limits on child labor, and rules requiring payment of overtime pay for those in jobs that are paid by the hour and exceed 40 hours per week</td>
</tr>
<tr>
<td>Taft-Hartley Act of 1947</td>
<td>Allows states to decide whether all workers at a firm can be required to join a union as a condition of employment; in the case of a disruptive union strike, permits the president to declare a “cooling-off period” during which workers have to return to work</td>
</tr>
<tr>
<td>Civil Rights Act of 1964</td>
<td>Title VII of the Act prohibits discrimination in employment on the basis of race, gender, national origin, religion, or sexual orientation</td>
</tr>
<tr>
<td>Occupational Health and Safety Act of 1970</td>
<td>Creates the Occupational Safety and Health Administration (OSHA), which protects workers from physical harm in the workplace</td>
</tr>
<tr>
<td>Employee Retirement and Income Security Act of 1974</td>
<td>Regulates employee pension rules and benefits</td>
</tr>
<tr>
<td>Pregnancy Discrimination Act of 1978</td>
<td>Prohibits discrimination against women in the workplace who are planning to get pregnant or who are returning to work after pregnancy</td>
</tr>
<tr>
<td>Immigration Reform and Control Act of 1986</td>
<td>Prohibits hiring of illegal immigrants; requires employers to ask for proof of citizenship; protects rights of legal immigrants</td>
</tr>
<tr>
<td>Worker Adjustment and Retraining Notification Act of 1988</td>
<td>Requires employers with more than 100 employees to provide written notice 60 days before plant closings or large layoffs</td>
</tr>
<tr>
<td>Americans with Disabilities Act of 1990</td>
<td>Prohibits discrimination against those with disabilities and requires reasonable accommodations for them on the job</td>
</tr>
<tr>
<td>Law</td>
<td>Protection</td>
</tr>
<tr>
<td>-------------------------------------------------------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Family and Medical Leave Act of 1993</td>
<td>Allows employees to take up to 12 weeks of unpaid leave per year for family</td>
</tr>
<tr>
<td></td>
<td>reasons, including birth or family illness</td>
</tr>
<tr>
<td>Pension Protection Act of 2006</td>
<td>Penalizes firms for underfunding their pension plans and gives employees</td>
</tr>
<tr>
<td></td>
<td>more information about their pension accounts</td>
</tr>
<tr>
<td>Lilly Ledbetter Fair Pay Act of 2009</td>
<td>Restores protection for pay discrimination claims on the basis of sex, race,</td>
</tr>
<tr>
<td></td>
<td>national origin, age, religion, or disability</td>
</tr>
</tbody>
</table>

There are two sources of imperfect competition in labor markets. These are demand side sources, that is, labor market power by employers, and supply side sources: labor market power by employees. Let’s begin by discussing demand side sources.

**Labor Market Power by Employers**

A competitive labor market is one where there are many potential employers for a given type of worker, say a secretary or an accountant. Suppose there is only one employer in a labor market. Because that employer has no direct competition in hiring, if they offer lower wages than would exist in a competitive market, employees will have few options. If they want a job, they must accept the offered wage rate. Since the employer is exploiting its market power, we call the firm a monopsony. The classical example of monopsony is the sole coal company in a West Virginia town. If coal miners want to work, they must accept what the coal company is paying. This is not the only example of monopsony. Think about surgical nurses in a town with only one hospital. Employers that have at least some market power over potential employees is not that unusual. After all, most firms have many employees while there is only one employer. Thus, even if there is some competition for workers, it may not feel that way to potential employees unless they do their research and find the opposite.

How does market power by an employer affect labor market outcomes? Intuitively, one might think that wages will be lower than in a competitive labor market. Let’s prove it. We will tell the story for a monopsonist, but the results will be qualitatively similar, although less extreme for any firm with labor market power.

Think back to monopoly. The good news is that because the monopolist is the sole supplier in the market, it can charge any price it wishes. The bad news is that if it wants to sell a greater quantity of output, it must lower the price it charges. Monopsony is analogous. Because the monopsonist is the sole employer in a labor market, it can offer any wage that it wishes. However, because they face the market supply curve for labor, if they want to hire more workers, they must raise the wage they pay. This creates a quandary, which we can understand by introducing a new concept: the marginal cost of labor. The marginal cost of labor is the cost to the firm of hiring one more worker. However, here is the thing: we assume that the firm is determining how many workers to hire in total. They are not hiring sequentially. Let’s look how this plays out with the example in Table 2.

<table>
<thead>
<tr>
<th>Supply of Labor</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wage Rate</td>
<td>$1 per hour</td>
<td>$2 per hour</td>
<td>$3 per hour</td>
<td>$4 per hour</td>
<td>$5 per hour</td>
</tr>
<tr>
<td>Total Cost of Labor</td>
<td>$1</td>
<td>$4</td>
<td>$9</td>
<td>$16</td>
<td>$25</td>
</tr>
<tr>
<td>Marginal Cost of Labor</td>
<td>$1</td>
<td>$3</td>
<td>$5</td>
<td>$7</td>
<td>$9</td>
</tr>
</tbody>
</table>

There are a couple of things to notice from the table. First, the marginal cost increases faster than the wage rate. In fact, the marginal cost of hiring an additional worker (beyond the first one) is always greater than the wage necessary to pay them. This is shown in Figure 1, where the MCe is above the supply curve of labor for any quantity of workers (again, beyond the first). This is because to hire one more worker requires paying a higher wage rate, not just for the new worker but for all the previous hires who would have accepted the lower wage. We can see this graphically in Figure 1.
Since monopsonies are the sole demander for labor, they face the market supply curve for labor. In order to increase employment they must raise the wage they pay not just for new workers, but for all the existing workers they could have hired at the previous lower wage. As a result, the marginal cost of additional hiring labor is greater than the wage, and thus for any level of employment (above the first worker), $MC_L$ is above the Market Supply of Labor.

If the firm wants to maximize profits, it will hire labor up to the point $L_m$ where $D_L = VMP$ (or MRP) = $MC_L$, as Figure 2 below shows. Then, the supply curve for labor shows the wage the firm will have to pay to attract $L_m$ workers. Graphically, we can draw a vertical line up from $L_m$ to the Supply Curve for labor and then read the wage $W_m$ off the vertical axis to the left.

A monopsony will hire workers up to the point $L_m$ where its demand for labor equals the marginal cost of additional labor, paying the wage $W_m$ given by the supply curve of labor necessary to obtain $L_m$ workers.
How does this outcome compare to what would occur in a perfectly competitive market? A competitive market would operate where $D_L = S_L$, hiring $L_C$ workers and paying $W_C$ wage. In other words, under monopsony employers hire fewer workers and pay a lower wage. While pure monopsony may be rare, many employers have some degree of market power in labor markets. The outcomes for those employers will be qualitatively similar though not as extreme as monopsony.

![Diagram](image.png)

**Figure 3. Comparison of labor market outcomes: Monopsony vs. Perfect Competition.** A monopsony hires fewer workers $L_m$ than would be hired in a competitive labor market $L_C$. In exploiting its market power, the monopsony can also pay a lower wage $W_m$ than workers would earn in a competitive labor market $W_C$.

**Try It**

Visit this page in your course online to check your understanding.

**Glossary**

- **marginal cost of labor**: the cost to the firm of hiring one more worker
- **monopsony**: a labor market where there is only one employer

**Licensing & Attributions**

CC licensed content, Original

- Wages and Employment in an Imperfectly Competitive Labor Market. **Authored by**: OpenStax College. **Located at**: https://cnx.org/contents/v6mOh_-pf84.44.2fbzMimj@11/Market-Power-on-the-Supply-Sid. **License**: CC BY. **Attribution**. **License Terms**: Download for free at http://cnx.org/contents/bc498e1f-efe9-43a0-8dea-d3569ad09a82@4.44

**LEARN BY DOING: LABOR MARKET POWER BY EMPLOYERS**
LEARN BY DOING: WORKERS AND WAGES

Try It

Play the simulation below multiple times to see how different choices lead to different outcomes. All simulations allow unlimited attempts so that you can gain experience applying the concepts.
Visit this page in your course online to use this simulation.

LABOR MARKET POWER BY EMPLOYEES

Learning Objectives

- Describe the role and impact of labor unions
- Explain how the existence of a labor union affects wages and employment in a labor market

Labor Market Power by Employees

A labor union is an organization of workers that negotiates with employers over wages and working conditions. A labor union seeks to change the balance of power between employers and workers by requiring employers to deal with
workers collectively, rather than as individuals. As such, a labor union operates like a monopoly in a labor market. We sometimes call negotiations between unions and firms collective bargaining.

The subject of labor unions can be controversial. Supporters of labor unions view them as the workers’ primary line of defense against efforts by profit-seeking firms to hold down wages and benefits. Critics of labor unions view them as having a tendency to grab as much as they can in the short term, even if it means injuring workers in the long run by driving firms into bankruptcy or by blocking the new technologies and production methods that lead to economic growth. We will start with some facts about union membership in the United States.

Facts about Union Membership and Pay

According to the U.S. Bureau of Labor and Statistics, about 10.7% of all U.S. workers belong to unions. Following are some facts about unions for 2016:

- 11.2% of U.S. male workers belong to unions; 10.2% of female workers do
- 10.5% of white workers, 13% of black workers, and 8.8% of Hispanic workers belong to unions
- 11.8% of full-time workers and 5.7% of part-time workers are union members
- 5.11% of workers ages 16–24 belong to unions, as do 13.9% of workers ages 45-54
- Occupations in which relatively high percentages of workers belong to unions are the federal government (27.4% belong to a union), state government (29.6%), local government (40.3%); transportation and utilities (15.1%); natural resources, construction, and maintenance (16.3%); and production, transportation, and material moving (13.7%)
- Occupations that have relatively low percentages of unionized workers are agricultural workers (1.3%), financial services (2.4%), professional and business services (2.4%), leisure and hospitality (2.7%), and wholesale and retail trade (4.2%)

In summary, the percentage of workers belonging to a union is higher for men than women; higher for blacks than for whites or Hispanics; higher for the 45–64 age range; and higher among workers in government and manufacturing than workers in agriculture or service-oriented jobs. Table 1 lists the largest U.S. labor unions and their membership.

<table>
<thead>
<tr>
<th>Union</th>
<th>Membership</th>
</tr>
</thead>
<tbody>
<tr>
<td>National Education Association (NEA)</td>
<td>2.9 million</td>
</tr>
<tr>
<td>Service Employees International Union (SEIU)</td>
<td>1.9 million</td>
</tr>
<tr>
<td>American Federation of Teachers (AFT)</td>
<td>1.5 million</td>
</tr>
<tr>
<td>International Brotherhood of Teamsters (IBT)</td>
<td>1.3 million</td>
</tr>
<tr>
<td>The American Federation of State, County, and Municipal Workers (AFSCME)</td>
<td>1.3 million</td>
</tr>
<tr>
<td>United Food and Commercial Workers International Union</td>
<td>1.3 million</td>
</tr>
<tr>
<td>International Brotherhood of Electrical Workers (IBEW)</td>
<td>662,000</td>
</tr>
<tr>
<td>United Steelworkers</td>
<td>591,000</td>
</tr>
<tr>
<td>International Association of Machinists and Aerospace Workers</td>
<td>569,000</td>
</tr>
<tr>
<td>International Union, United Automobile, Aerospace and Agricultural Implement Workers of America (UAW)</td>
<td>408,000</td>
</tr>
</tbody>
</table>

In terms of pay, benefits, and hiring, U.S. unions offer a good news/bad news story. The good news for unions and their members is that their members earn about 20% more than nonunion workers, even after adjusting for factors such as years of work experience and education level. The bad news for unions is that the share of U.S. workers who belong to a labor union has been steadily declining for 50 years, as Figure 1 shows. About one-quarter of all U.S. workers belonged to a union in the mid-1950s, but only 11.1% of U.S. workers are union members today. If you leave out
government workers (which includes teachers in public schools), only 6.6% of the workers employed by private firms now work for a union.

Figure 1. Percentage of Wage and Salary Workers Who Are Union Members. The share of wage and salary workers who belong to unions rose sharply in the 1930s and 1940s, but has tailed off since then to 10.7% of all workers in 2016.

Higher Wages for Union Workers

How does a union affect wages and employment? Because a union is the sole supplier of labor, it can act like a monopoly and ask for whatever wage rate it can obtain for its workers. If employers need workers, they have to meet the union’s wage demand.

What are the limits on how much higher pay union workers can receive? To analyze these questions, let’s consider a situation where all firms in an industry must negotiate with a single union, and no firm is allowed to hire nonunion labor. If no labor union existed in this market, then equilibrium (E) in the labor market would occur at the intersection of the demand for labor (D) and the supply of labor (S) as we see in Figure 2. The union can, however, threaten that, unless firms agree to the wages they demand, the workers will strike. As a result, the labor union manages to achieve, through negotiations with the firms, a union wage of Wu for its members, above what the equilibrium wage would otherwise have been.
Without a union, the equilibrium at \( E \) would have involved the wage \( W_e \) and the quantity of labor \( Q_e \). However, the union is able to use its bargaining power to raise the wage to \( W_u \). The result is an excess supply of labor for union jobs. That is, a quantity of labor supplied, \( Q_s \), is greater than firms' quantity demanded for labor, \( Q_d \).

This labor market situation resembles what a monopoly firm does in selling a product, but in this case a union is a monopoly selling labor to firms. At the higher union wage \( W_u \), the firms in this industry will hire less labor than they would have hired in equilibrium. Moreover, an excess supply of workers want union jobs, but firms will not be hiring for such jobs.

From the union point of view, workers who receive higher wages are better off. However, notice that the quantity of workers (\( Q_d \)) hired at the union wage \( W_u \) is smaller than the quantity \( Q_e \) that the firm would have hired at the original equilibrium wage. A sensible union must recognize that when it pushes up the wage, it also reduces the firms' incentive to hire. This situation does not necessarily mean that union workers are fired. Instead, it may be that when union workers move on to other jobs or retire, they are not always replaced, or perhaps when a firm expands production, it expands employment somewhat less with a higher union wage than it would have done with the lower equilibrium wage. Other situations could be that a firm decides to purchase inputs from nonunion producers, rather than producing them with its own highly paid unionized workers, or perhaps the firm moves or opens a new facility in a state or country where unions are less powerful.

From the firm's point of view, the key question is whether union workers' higher wages are matched by higher productivity. If so, then the firm can afford to pay the higher union wages and, the demand curve for "unionized" labor could actually shift to the right. This could reduce the job losses as the equilibrium employment level shifts to the right and the difference between the equilibrium and the union wages will have been reduced. If worker unionization does not increase productivity, then the higher union wage will cause lower profits or losses for the firm.

Union workers might have higher productivity than nonunion workers for a number of reasons. First, higher wages may elicit higher productivity. Second, union workers tend to stay longer at a given job, a trend that reduces the employer's costs for training and hiring and results in workers with more years of experience. Many unions also offer job training and apprenticeship programs.

In addition, firms that are confronted with union demands for higher wages may choose production methods that involve more physical capital and less labor, resulting in increased labor productivity. Table 2 provides an example. Assume that a firm can produce a home exercise cycle with three different combinations of labor and manufacturing equipment. Say that the firm pays labor $16 an hour (including benefits) and the machines for manufacturing cost $200 each. Under these circumstances, the total cost of producing a home exercise cycle will be lowest if the firm adopts the plan of 50 hours of labor and one machine, as the table shows ($16 x 50 hours is $800, plus the $200 manufacturing costs for a total of $1000). Now, suppose that a union negotiates a wage of $20 an hour including benefits. In this case, it makes no difference to the firm whether it uses more hours of labor and fewer machines or less labor and more machines, although it might prefer to use more machines and to hire fewer union workers. (After all, machines never threaten to strike—but they do not buy the final product or service either.)
Table 2. Three Production Choices to Manufacture a Home Exercise Cycle

<table>
<thead>
<tr>
<th>Hours of Labor</th>
<th>Number of Machines</th>
<th>Cost of Labor + Cost of Machine $16/hour</th>
<th>Cost of Labor + Cost of Machine $20/hour</th>
<th>Cost of Labor + Cost of Machine $24/hour</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>3</td>
<td>$480 + $600 = $1,080</td>
<td>$600 + $600 = $1,200</td>
<td>$720 + $600 = $1,320</td>
</tr>
<tr>
<td>40</td>
<td>2</td>
<td>$640 + $400 = $1,040</td>
<td>$800 + $400 = $1,200</td>
<td>$960 + $400 = $1,360</td>
</tr>
<tr>
<td>50</td>
<td>1</td>
<td>$800 + $200 = $1,000</td>
<td>$1,000 + $200 = $1,200</td>
<td>$1,200 + $200 = $1,400</td>
</tr>
</tbody>
</table>

In the final column of the table, the wage has risen to $24 an hour. In this case, the firm clearly has an incentive for using the plan that involves paying for fewer hours of labor and using three machines. If management responds to union demands for higher wages by investing more in machinery, then union workers can be more productive because they are working with more or better physical capital equipment than the typical nonunion worker. However, the firm will need to hire fewer workers.

In some cases, unions have discouraged the use of labor-saving physical capital equipment—out of the reasonable fear that new machinery will reduce the number of union jobs. For example, in 2002, the union representing longshoremen who unload ships and the firms that operate shipping companies and port facilities staged a work stoppage that shut down the ports on the western coast of the United States. Two key issues in the dispute were the desire of the shipping companies and port operators to use handheld scanners for record-keeping and computer-operated cabs for loading and unloading ships—changes which the union opposed, along with overtime pay. President Obama threatened to use the Labor Management Relations Act of 1947—commonly known as the Taft-Hartley Act—where a court can impose an 80-day “cooling-off period” in order to allow time for negotiations to proceed without the threat of a work stoppage. Federal mediators were called in, and the two sides agreed to a deal in February 2015. The ultimate agreement allowed the new technologies, but also kept wages, health, and pension benefits high for workers. In the past, presidential use of the Taft-Hartley Act sometimes has made labor negotiations more bitter and argumentative but, in this case, it seems to have smoothed the road to an agreement.

In other instances, unions have proved quite willing to adopt new technologies. In one prominent example, during the 1950s and 1960s, the United Mineworkers union demanded that mining companies install labor-saving machinery in the mines. The mineworkers’ union realized that over time, the new machines would reduce the number of jobs in the mines, but the union leaders also knew that the mine owners would have to pay higher wages if the workers became more productive, and mechanization was a necessary step toward greater productivity.

In fact, in some cases union workers may be more willing to accept new technology than nonunion workers, because the union workers believe that the union will negotiate to protect their jobs and wages, whereas nonunion workers may be more concerned that the new technology will replace their jobs. In addition, union workers, who typically have higher job market experience and training, are likely to suffer less and benefit more than non-union workers from the introduction of new technology. Overall, it is hard to make a definitive case that union workers as a group are always either more or less welcoming to new technology than are nonunion workers.

The Decline in U.S. Union Membership

The proportion of U.S. workers belonging to unions has declined dramatically since the early 1950s. Economists have offered a number of possible explanations:

- The shift from manufacturing to service industries
- The force of globalization and increased competition from foreign producers
- A reduced desire for unions because of the workplace protection laws now in place
- U.S. legal environment that makes it relatively more difficult for unions to organize workers and expand their membership

Let's discuss each of these four explanations in more detail.

A first possible explanation for the decline in the share of U.S. workers belonging to unions involves the patterns of job growth in the manufacturing and service sectors of the economy as Figure 3 shows. The U.S. economy had about 15 million manufacturing jobs in 1960. This total rose to 19 million by the late 1970s and then declined to 17 million in 2013. Meanwhile, the number of jobs in service industries and in government combined rose from 35 million in 1960 to over
118 million by 2013, according to the Bureau of Labor Statistics. Because over time unions were stronger in manufacturing than in service industries, the growth in jobs was not happening where the unions were. It is interesting to note that government workers comprise several of the biggest unions in the country, including the American Federation of State, County and Municipal Employees (AFSCME); the Service Employees International Union; and the National Education Association. Table 1 lists the membership of each of these unions. Outside of government employees, however, unions have not had great success in organizing the service sector.

![Figure 3. The Growth Service Jobs.](image)

Jobs in services have increased dramatically in the last few decades. Jobs in government have increased modestly until 1990 and then declined slightly since then. Jobs in manufacturing peaked in the late 1970s and have declined more than a third since then.

A second explanation for the decline in the share of unionized workers looks at import competition. Starting in the 1960s, U.S. carmakers and steelmakers faced increasing competition from Japanese and European manufacturers. As sales of imported cars and steel rose, the number of jobs in U.S. auto manufacturing fell. This industry is heavily unionized. Not surprisingly, membership in the United Auto Workers, which was 975,000 in 1985, had fallen to roughly 390,000 by 2015. Import competition not only decreases the employment in sectors where unions were once strong, but also decreases the bargaining power of unions in those sectors. However, as we have seen, unions that organize public-sector workers, who are not threatened by import competition, have continued to see growth.

A third possible reason for the decline in the number of union workers is that citizens often call on their elected representatives to pass laws concerning work conditions, overtime, parental leave, regulation of pensions, and other issues. Unions offered strong political support for these laws aimed at protecting workers but, in an ironic twist, the passage of those laws then made many workers feel less need for unions.

These first three possible reasons for the decline of unions are all somewhat plausible, but they have a common problem. Most other developed economies have experienced similar economic and political trends, such as the shift from manufacturing to services, globalization, and increasing government social benefits and regulation of the workplace. Clearly there are cultural differences between countries as to their acceptance of unions in the workplace. The share of the population belonging to unions in other countries is very high compared with the share in the United States. Table 3 shows the proportion of workers in a number of the world's high-income economies who belong to unions. The United States is near the bottom, along with France and Spain. The last column shows union coverage, defined as including those workers whose wages are determined by a union negotiation even if the workers do not officially belong to the union. In the United States, union membership is almost identical to union coverage. However, in many countries, the wages of many workers who do not officially belong to a union are still determined by collective bargaining between unions and firms.

Table 3. International Comparisons of Union Membership and Coverage in 2012 (Source, CIA World Factbook, retrieved from www.cia.gov)
<table>
<thead>
<tr>
<th>Country</th>
<th>Union Density: Percentage of Workers Belonging to a Union</th>
<th>Union Coverage: Percentage of Workers Whose Wages Are Determined by Union Bargaining</th>
</tr>
</thead>
<tbody>
<tr>
<td>Austria</td>
<td>37%</td>
<td>99%</td>
</tr>
<tr>
<td>France</td>
<td>9%</td>
<td>95%</td>
</tr>
<tr>
<td>Germany</td>
<td>26%</td>
<td>63%</td>
</tr>
<tr>
<td>Japan</td>
<td>22%</td>
<td>23%</td>
</tr>
<tr>
<td>Netherlands</td>
<td>25%</td>
<td>82%</td>
</tr>
<tr>
<td>Spain</td>
<td>11.3%</td>
<td>81%</td>
</tr>
<tr>
<td>Sweden</td>
<td>82%</td>
<td>92%</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>29%</td>
<td>35%</td>
</tr>
<tr>
<td>United States</td>
<td>11.1%</td>
<td>12.5%</td>
</tr>
</tbody>
</table>

These international differences in union membership suggest a fourth reason for the decline of union membership in the United States: perhaps U.S. laws are less friendly to the formation of unions than such laws in other countries. The close connection between union membership and a friendly legal environment is apparent in the history of U.S. unions. The great rise in union membership in the 1930s followed the passage of the National Labor-Management Relations Act of 1935, which specified that workers had a right to organize unions and that management had to give them a fair chance to do so. The U.S. government strongly encouraged forming unions during the early 1940s in the belief that unions would help to coordinate the all-out production efforts needed during World War II. However, after World War II came the passage of the Taft-Hartley Act of 1947, which gave states the power to allow workers to opt out of the union in their workplace if they so desired. This law made the legal climate less encouraging to those seeking to form unions, and union membership levels soon started declining.

The procedures for forming a union differ substantially from country to country. For example, the procedures in the United States and those in Canada are strikingly different. When a group of workers wish to form a union in the United States, they announce this fact and set an election date when the firm’s employees will vote in a secret ballot on whether to form a union. Supporters of the union lobby for a “yes” vote, and the firm’s management lobbies for a “no” vote—often even hiring outside consultants for assistance in swaying workers to vote “no.” In Canada, by contrast, a union is formed when a sufficient proportion of workers (usually about 60%) sign an official card saying that they want a union. There is no separate “election date.” The management of Canadian firms is limited by law in its ability to lobby against the union. In addition, although it is illegal to discriminate and fire workers based on their union activity in the United States, the penalties are slight, making this a not so costly way of deterring union activity. In short, forming unions is easier in Canada—and in many other countries—than in the United States.

In summary, union membership in the United States is lower than in many other high-income countries, a difference that may be due to different legal environments and cultural attitudes toward unions.

Bilateral Monopolies

What happens when there is market power on both sides of the labor market, in other words, when a union meets a monopsony? Economists call such a situation a bilateral monopoly.
Figure 4. Bilateral Monopoly. Employment, $L^*$, will be lower in a bilateral monopoly than in a competitive labor market, but the equilibrium wage is indeterminate, somewhere in the range between $W_u$, what the union would choose, and $W_m$, what the monopsony would choose.

Figure 4 shows the graph for bilateral monopoly, which is a combination of the graphs for monopsonies and unions. A monopsony wants to reduce wages as well as employment, $W_m$ and $L^*$ in the figure. A union wants to increase wages, but at the cost of lower employment, $W_u$ and $L^*$ in the figure. Since both sides want to reduce employment, we can be sure that the outcome will be lower employment compared to a competitive labor market. What happens to the wage, though, is based on the monopsony’s relative bargaining power compared to the union. The actual outcome is indeterminate in the graph, but it will be closer to $W_u$ if the union has more power and closer to $W_m$ if the monopsonist has more power.

Try It

Visit this page in your course online to check your understanding.

Try It

These questions allow you to get as much practice as you need, as you can click the link at the top of the first question (“Try another version of these questions”) to get a new set of questions. Practice until you feel comfortable doing the questions.

Visit this page in your course online to practice before taking the quiz.

Glossary

**bilateral monopoly** a labor market with a monopsony on the demand side and a union on the supply side

**monopsony** a labor market where there is only one employer
INTRODUCTION TO DISCRIMINATION IN LABOR MARKETS

What you’ll learn to do: analyze the economic implications of discrimination and immigration policies

Figure 1. Policies and Wages. Much controversy surrounds government policies that would dissuade workplace discrimination.

Discrimination occurs in a labor market when employers pay workers with the same economic characteristics, such as education, experience, and skill, are paid different amounts because of race, gender, religion, age, or disability status. In the United States, female workers on average earn less than male workers, and black workers on average earn less than white workers. There is controversy over the extent to which pay differences are caused by differences in factors like education and job experience, or by pure discrimination.

Free markets can allow discrimination to occur, but the threat of a loss of sales or a loss of productive workers can also create incentives for a firm not to discriminate. A range of public policies can be used to reduce earnings gaps between men and women or between white and other racial/ethnic groups: requiring equal pay for equal work, and attaining more equal educational outcomes.

In this section, we will examine these policies as well as the economic take on immigration.

EMPLOYMENT DISCRIMINATION
Learning Objectives

- Analyze earnings gaps based on race and gender
- Explain the impact of discrimination in a competitive market

Discrimination involves acting on the belief that members of a certain group are inferior solely because of a factor such as race, gender, or religion. There are many types of discrimination but the focus here will be on discrimination in labor markets, which arises if workers with the same skill levels—as measured by education, experience, and expertise—receive different pay or have different job opportunities because of their race or gender.

Earnings Gaps by Race and Gender

A possible signal of labor market discrimination is when an employer pays one group less than another. Figure 1 shows the average wage of black workers as a ratio of the average wage of white workers and the average wage of female workers as a ratio of the average wage of male workers. Research by the economists Francine Blau and Laurence Kahn shows that the gap between the earnings of women and men did not move much in the 1970s, but has declined since the 1980s. According to the U.S. Census, the gap between the earnings of blacks and whites diminished in the 1970s, but has not changed in 50 years. In both gender and race, an earnings gap remains.

An earnings gap between average wages, in and of itself, does not prove that discrimination is occurring in the labor market. We need to apply the same productivity characteristics to all parties (employees) involved. Gender discrimination in the labor market occurs when employers pay women less than men despite having comparable levels of education, experience, and expertise. (Read the following feature about the sex-discrimination suit brought against Wal-Mart.) Similarly, racial discrimination in the labor market exists when employers pay racially diverse employees less than their coworkers of the majority race despite having comparable levels of education, experience, and expertise. To bring a successful gender discrimination lawsuit, a female employee must prove the employer is paying her less than a male employee who holds a similar job, with similar educational attainment, and with similar expertise. Likewise, someone who wants to sue on the grounds of racial discrimination must prove that the employer pays him or her less...
than an employee of another race who holds a similar job, with similar educational attainment, and with similar expertise.

What was the sex-discrimination case against Wal-Mart?

In one of the largest class-action sex-discrimination cases in U.S. history, 1.2 million female employees of Wal-Mart claimed that the company engaged in wage and promotion discrimination. In 2011, the Supreme Court threw out the case on the grounds that the group was too large and too diverse to consider the case a class action suit. Lawyers for the women regrouped and are now suing in smaller groups. Part of the difficulty for the female employees is that the court said that local managers made pay and promotion decisions that were not necessarily the company's policies as a whole. Consequently, female Wal-Mart employees in Texas are arguing that their new suit will challenge the management of a “discrete group of regional district and store managers.” They claim these managers made biased pay and promotion decisions. However, in 2013, a federal district court rejected a smaller California class action suit against the company.

On other issues, Wal-Mart made the news again in 2013 when the National Labor Relations Board found Wal-Mart guilty of illegally penalizing and firing workers who took part in labor protests and strikes. Wal-Mart has already paid $11.7 million in back wages and compensation damages to women in Kentucky who were denied jobs due to their sex.

Investigating the Female/Male Earnings Gap

As a result of changes in law and culture, women began to enter the paid workforce in substantial numbers in the mid-to late-twentieth century. By 2014, 58.1% of adult women held jobs while 72.0% of adult men did. Moreover, along with entering the workforce, women began to ratchet up their education levels. In 1971, 44% of undergraduate college degrees went to women. By 2014, women received 56% of bachelor's degrees. In 1970, women received 5.4% of the degrees from law schools and 8.4% of the degrees from medical schools. By 2014, women were receiving 47% of the law degrees and 48.0% of the medical degrees. These gains in education and experience have reduced the female/male wage gap over time. However, concerns remain about the extent to which women have not yet assumed a substantial share of the positions at the top of the largest companies or in the U.S. Congress.

There are factors that can lower women's average wages. Women are likely to bear a disproportionately large share of household responsibilities. A mother of young children is more likely to drop out of the labor force for several years or work on a reduced schedule than is the father. As a result, women in their 30s and 40s are likely, on average, to have less job experience than men. In the United States, childless women with the same education and experience levels as men are typically paid comparably. However, women with families and children are typically paid about 7% to 14% less than other women of similar education and work experience. (Meanwhile, married men earn about 10% to 15% more than single men with comparable education and work experience.)

We possibly could call the different patterns of family responsibilities discrimination, but it is primarily rooted in America’s social patterns of discrimination, which involve the roles that fathers and mothers play in child-rearing, rather than discrimination by employers in hiring and salary decisions.

Watch It

This video examines some of the causes of the wage gap.
Watch this video online: https://youtu.be/mpE8ttCEd-w

Investigating the Black/White Earnings Gap

Blacks experienced blatant labor market discrimination during much of the twentieth century. Until the passage of the Civil Rights Act of 1964, it was legal in many states to refuse to hire a black worker, regardless of the credentials or experience of that worker. Moreover, blacks were often denied access to educational opportunities, which in turn meant that they had lower levels of qualifications for many jobs. At least one economic study has shown that the 1964 law is partially responsible for the narrowing of the gap in black–white earnings in the late 1960s and into the 1970s. For
example, the ratio of total earnings of black male workers to white male workers rose from 62% in 1964 to 75.3% in 2013, according to the Bureau of Labor Statistics.

<table>
<thead>
<tr>
<th>Table 1. Educational Attainment by Race and Ethnicity in 2015 (Source: <a href="http://www.census.gov/hhes/socdemo/education/data/cps/2014/tables.html">http://www.census.gov/hhes/socdemo/education/data/cps/2014/tables.html</a>)</th>
</tr>
</thead>
<tbody>
<tr>
<td>White</td>
</tr>
<tr>
<td>-------------------</td>
</tr>
<tr>
<td>Completed four years of high school or more</td>
</tr>
<tr>
<td>Completed four years of college or more</td>
</tr>
</tbody>
</table>

However, the earnings gap between black and white workers has not changed as much as the earnings gap between men and women has in the last half century. The remaining racial gap seems related both to continuing differences in education levels and to the presence of discrimination. Table 1 shows that the percentage of blacks who complete a four-year college degree remains substantially lower than the percentage of whites who complete college. According to the U.S. Census, both whites and blacks have higher levels of educational attainment than Hispanics and lower levels than Asians. The lower average levels of education for black workers surely explain part of the earnings gap. In fact, black women who have the same levels of education and experience as white women receive, on average, about the same level of pay. One study shows that white and black college graduates have identical salaries immediately after college; however, the racial wage gap widens over time, an outcome that suggests the possibility of continuing discrimination. Another study conducted a field experiment by responding to job advertisements with fictitious resumes with either very African American sounding names or very white sounding names and found out that white names received 50 percent more callbacks for interviews. This is suggestive of discrimination in job opportunities. Further, as the following feature explains, there is evidence to support that discrimination in the housing market is connected to employment discrimination.

### How is discrimination in the housing market connected to employment discrimination?

In a recent study by the U.S. Department of Housing and Urban Development (HUD) department, realtors showed black homebuyers 18 percent fewer homes compared to white homebuyers. Realtors showed Asians 19 percent fewer properties. Additionally, Hispanics experienced more discrimination in renting apartments and underwent stiffer credit checks than white renters. In a 2012 study by HUD and the nonprofit Urban Institute, Hispanic testers who contacted agents about advertised rental units received information about 12 percent fewer units available and were shown seven percent fewer units than white renters. The $9 million study, based on research in 28 metropolitan areas, concluded that blatant “door slamming” forms of discrimination are on the decline but that the discrimination that does exist is harder to detect, and as a result, more difficult to remedy. According to the Chicago Tribune, HUD Secretary Shaun Donovan, who served in his role from 2009-2014, told reporters, “Just because it’s taken on a hidden form doesn’t make it any less harmful. You might not be able to move into that community with the good schools.”

The lower levels of education for black workers can also be a result of discrimination—although it may be pre-labor market discrimination, rather than direct discrimination by employers in the labor market. For example, if discrimination in housing markets causes black families to live clustered together in certain poorer neighborhoods, then the black children will continue to have lower educational attainment than their white counterparts and, consequently, not be able to obtain the higher paying jobs that require higher levels of education. Another element to consider is that in the past, when blacks were effectively barred from many high-paying jobs, obtaining additional education could have seemed somewhat pointless, because the educational degrees would not pay off. Even though the government has legally abolished labor market discrimination, it can take some time to establish a culture and a tradition of valuing education highly. Additionally, a legacy of past discrimination may contribute to an attitude that blacks will have a difficult time succeeding in academic subjects. In any case, the impact of social discrimination in labor markets is more complicated than seeking to punish a few bigoted employers.

---

**Try It**

Visit this page in your course online to view this presentation.
Competitive Markets and Discrimination

Gary Becker (b. 1930), who won the Nobel Prize in economics in 1992, was one of the first to analyze discrimination in economic terms. Becker pointed out that while competitive markets can allow some employers to practice discrimination, it can also provide profit-seeking firms with incentives not to discriminate. Given these incentives, Becker explored the question of why discrimination persists.

If a business is located in an area with a large minority population and refuses to sell to minorities, it will cut into its own profits. If some businesses run by bigoted employers refuse to pay women and/or minorities a wage based on their productivity, then other profit-seeking employers can hire these workers. In a competitive market, if the business owners care more about the color of money than about the color of skin, they will have an incentive to make buying, selling, hiring, and promotion decisions strictly based on economic factors.

Do not underestimate the power of markets to offer at least a degree of freedom to oppressed groups. In many countries, cohesive minority groups like Jews and emigrant Chinese have managed to carve out a space for themselves through their economic activities, despite legal and social discrimination against them. Many immigrants, including those who come to the United States, have taken advantage of economic freedom to make new lives for themselves. However, history teaches that market forces alone are unlikely to eliminate discrimination. After all, discrimination against African Americans persisted in the market-oriented U.S. economy during the century between President Abraham Lincoln's Emancipation Proclamation, which freed the slaves in 1863, and the passage of the Civil Rights Act of 1964—and has continued since then, too.

Therefore, why does discrimination persist in competitive markets? Gary Becker sought to explain this persistence. Discriminatory impulses can emerge at a number of levels: among managers, among workers, and among customers. Consider the situation of a manager who is not personally prejudiced, but who has many workers or customers who are prejudiced. If that manager treats minority groups or women fairly, the manager may find it hurts the morale of prejudiced co-workers or drives away prejudiced customers. In such a situation, a policy of nondiscrimination could reduce the firm’s profits. After all, a business firm is part of society, and a firm that does not follow the societal norms is likely to suffer. Market forces alone are unlikely to overwhelm strong social attitudes about discrimination.

Public Policies to Reduce Discrimination

A first public policy step against discrimination in the labor market is to make it illegal. For example, the Equal Pay Act of 1963 said that employers must pay men and women who do equal work the same. The Civil Rights Act of 1964 prohibits employment discrimination based on race, color, religion, sex, or national origin. The Age Discrimination in Employment Act of 1967 prohibited discrimination on the basis of age against individuals who are 40 years of age or older. The Civil Rights Act of 1991 provides monetary damages in cases of intentional employment discrimination. The Pregnancy Discrimination Act of 1978 was aimed at prohibiting discrimination against women in the workplace who are planning to get pregnant, are pregnant, or are returning after pregnancy. Passing a law, however, is only part of the answer, since discrimination by prejudiced employers may be less important than broader social patterns.

These laws against discrimination have reduced the gender wage gap. A 2007 Department of Labor study compared salaries of men and women who have similar educational achievement, work experience, and occupation and found that the gender wage gap is only 5%.

In the case of the earnings gap between blacks and whites (and also between Hispanics and whites), probably the single largest step that could be taken at this point in U.S. history to close the earnings gap would be to reduce the gap in educational achievement. Part of the answer to this issue involves finding ways to improve the performance of schools, which is a highly controversial topic in itself. In addition, the education gap is unlikely to close unless black and Hispanic families and peer groups strengthen their culture of support for educational achievement.

Affirmative action is the name given to active efforts by government or businesses that give special rights to minorities in hiring and promotion to make up for past discrimination. Affirmative action, in its limited and not especially controversial form, means making an effort to reach out to a broader range of minority candidates for jobs. In its more aggressive and controversial form, affirmative action required government and companies to hire a specific number or percentage of minority employees. However, the U.S. Supreme Court has ruled against state affirmative action laws. Today, the government applies affirmative action policies only to federal contractors who have lost a discrimination lawsuit. The federal Equal Employment Opportunity Commission (EEOC) enforces this type of redress.
An Increasingly Diverse Workforce

Racial and ethnic diversity is on the rise in the U.S. population and work force. As Figure 2 shows, while the white Americans comprised 78% of the population in 2012, the U.S. Bureau of the Census projects that whites will comprise 69% of the U.S. population by 2060. Forecasters predict that the proportion of U.S. citizens who are of Hispanic background to rise substantially. Moreover, in addition to expected changes in the population, workforce diversity is increasing as the women who entered the workforce in the 1970s and 1980s are now moving up the promotion ladders within their organizations.

![Figure 2. Projected Changes in America's Racial and Ethnic Diversity.](image)

Figure 2. Projected Changes in America’s Racial and Ethnic Diversity. This figure shows projected changes in the ethnic makeup of the U.S. population by 2060. Note that “NHPI” stands for Native Hawaiian and Other Pacific Islander. “AIAN” stands for American Indian and Alaska Native. Source: US Department of Commerce

Regarding the future, optimists argue that the growing proportions of minority workers will break down remaining discriminatory barriers. The economy will benefit as an increasing proportion of workers from traditionally disadvantaged groups have a greater opportunity to fulfill their potential. Pessimists worry that the social tensions between men and women and between ethnic groups will rise and that workers will be less productive as a result. Anti-discrimination policy, at its best, seeks to help society move toward the more optimistic outcome.

**Link to Learning**

Take a look at the FRED database to find data on employment, foreign and native born civilian population, and labor force.

**Try It**

Visit this page in your course online to check your understanding.

**Glossary**

affirmative action: active efforts by government or businesses that give special rights to minorities in hiring, promotion, or access to education to make up for past discrimination

discrimination: actions based on the belief that members of a certain group or groups are in some way inferior solely because of a factor such as race, gender, or religion
IMMIGRATION

Learning Objectives

- Discuss the economic implications of immigration

Most Americans would be outraged if a law prevented them from moving to another city or another state. However, when the conversation turns to crossing national borders and are about other people arriving in the United States, laws preventing such movement often seem more reasonable. Some of the tensions over immigration stem from worries over how it might affect a country’s culture, including differences in language, and patterns of family, authority, or gender relationships. Economics does not have much to say about such cultural issues. Some of the worries about immigration do, however, have to do with its effects on wages and income levels, and how it affects government taxes and spending. On those topics, economists have insights and research to offer.

Historical Patterns of Immigration

Supporters and opponents of immigration look at the same data and see different patterns. Those who express concern about immigration levels to the United States point to graphics like Figure 1 which shows total inflows of immigrants decade by decade through the twentieth century. Clearly, the level of immigration has been high and rising in recent years, reaching and exceeding the towering levels of the early twentieth century. However, those who are less worried about immigration point out that the high immigration levels of the early twentieth century happened when total population was much lower. Since the U.S. population roughly tripled during the twentieth century, the seemingly high levels in immigration in the 1990s and 2000s look relatively smaller when they are divided by the population.
From where have the immigrants come? Immigrants from Europe were more than 90% of the total in the first decade of the twentieth century, but less than 20% of the total by the end of the century. By the 2000s, about half of U.S. immigration came from the rest of the Americas, especially Mexico, and about a quarter came from various countries in Asia.

Economic Effects of Immigration

A surge of immigration can affect the economy in a number of different ways. In this section, we will consider how immigrants might benefit the rest of the economy, how they might affect wage levels, and how they might affect government spending at the federal and local level.

To understand the economic consequences of immigration, consider the following scenario. Imagine that the immigrants entering the United States matched the existing U.S. population in age range, education, skill levels, family size, and occupations. How would immigration of this type affect the rest of the U.S. economy? Immigrants themselves would be much better off, because their standard of living would be higher in the United States. Immigrants would contribute to both increased production and increased consumption. Given enough time for adjustment, the range of jobs performed, income earned, taxes paid, and public services needed would not be much affected by this kind of immigration. It would be as if the population simply increased a little.

Now, consider the reality of recent immigration to the United States. Immigrants are not identical to the rest of the U.S. population. About one-third of immigrants over the age of 25 lack a high school diploma. As a result, many of the recent immigrants end up in jobs like restaurant and hotel work, lawn care, and janitorial work. This kind of immigration represents a shift to the right in the supply of unskilled labor for a number of jobs, which will lead to lower wages for these jobs. The middle- and upper-income households that purchase the services of these unskilled workers will benefit from these lower wages. However, low-skilled U.S. workers who must compete with low-skilled immigrants for jobs will tend to suffer from immigration.

The difficult policy questions about immigration are not so much about the overall gains to the rest of the economy, which seem to be real but small in the context of the U.S. economy, as they are about the disruptive effects of immigration in specific labor markets. One disruptive effect, as we noted, is that immigration weighted toward low-skill workers tends to reduce wages for domestic low-skill workers. A study by Michael S. Clune found that for each 10% rise in the number of employed immigrants with no more than a high school diploma in the labor market, high school students reduced their annual number of hours worked by 3%. The effects on wages of low-skill workers are not large—
perhaps in the range of decline of about 1%. These effects are likely kept low, in part, because of the legal floor of federal and state minimum wage laws. In addition, immigrants are also thought to contribute to increased demand for local goods and services which can stimulate the local low skilled labor market. It is also possible that employers, in the face of abundant low-skill workers may choose production processes which are more labor intensive than otherwise would have been. These various factors would explain the small negative wage effect that the native low-skill workers observed as a result of immigration.

Another potential disruptive effect is the impact on state and local government budgets. Many of the costs imposed by immigrants are costs that arise in state-run programs, like the cost of public schooling and of welfare benefits. However, many of the taxes that immigrants pay are federal taxes like income taxes and Social Security taxes. Many immigrants do not own property (such as homes and cars), so they do not pay property taxes, which are one of the main sources of state and local tax revenue. However, they do pay sales taxes, which are state and local, and the landlords of property they rent pay property taxes. According to the nonprofit Rand Corporation, the effects of immigration on taxes are generally positive at the federal level, but they are negative at the state and local levels in places where there are many low-skilled immigrants.

Proposals for Immigration Reform

The Congressional Jordan Commission of the 1990s proposed reducing overall levels of immigration and refocusing U.S. immigration policy to give priority to immigrants with higher skill levels. In the labor market, focusing on high-skilled immigrants would help prevent any negative effects on low-skilled workers’ wages. For government budgets, higher-skilled workers find jobs more quickly, earn higher wages, and pay more in taxes. Several other immigration-friendly countries, notably Canada and Australia, have immigration systems where those with high levels of education or job skills have a much better chance of obtaining permission to immigrate. For the United States, high tech companies regularly ask for a more lenient immigration policy to admit a greater quantity of highly skilled workers under the H1B visa program.

The Obama Administration proposed the so-called “DREAM Act” legislation, which would have offered a path to citizenship for illegal immigrants brought to the United States before the age of 16. Despite bipartisan support, the legislation failed to pass at the federal level. However, some state legislatures, such as California, have passed their own Dream Acts.

Between its plans for a border wall, increased deportation of undocumented immigrants, and even reductions in the number of highly skilled legal H1B immigrants, the Trump Administration has a much less positive approach to immigration. Most economists, whether conservative or liberal, believe that while immigration harms some domestic workers, the benefits to the nation exceed the costs. However, given the Trump Administration’s opposition, any significant immigration reform is likely on hold.

Watch It

Watch this video for a comprehensive overview on the economics of immigration.
Watch this video online: https://youtu.be/4XQXiCLzyAw
PUTTING IT TOGETHER: LABOR MARKETS

In this module, we learned about labor markets, wages, and other factors affecting employment, such as discrimination. Teachers and nurses are paid less than professional athletes because the market values the former less than the latter. In other words, our actions say that we are willing to pay professional athletes more than teachers and nurses. This may be because athletes are employed through the private sector while most teachers and nurses are employed by the public sector where the lack of market forces makes it harder for workers to be paid what they’re worth. Either way, it’s a statement about social values.

Urban sanitation engineers (i.e. garbage truck workers) get paid a decent wage, not because of the skills required for the job, but rather because of the difficult working conditions in summer and winter. Less “desirable” jobs have to pay more to attract workers.

Unionized workers earn more than non-union workers because unions are able to take advantage of monopoly power in the labor market. Just as a monopoly in the output market can charge a higher price than would be charged if the market were competitive, so unions can charge a higher wage.

The Increasing Value of a College Degree

At the beginning of the module, we discussed how the cost of college has increased dramatically in recent decades, causing many college students to take student loans to afford it. Despite this, the value of a college degree has never been higher. How can we explain this?

We can estimate the value of a bachelor’s degree as the difference in lifetime earnings between the average holder of a bachelor’s degree and the average high school graduate. This difference can be nearly $1 million. College graduates also have a significantly lower unemployment rate than those with lower educational attainments.

While a college degree holder’s wages have increased somewhat, the major reason for the increase in value of a bachelor’s degree has been the plummeting value of a high school diploma. In the twenty-first century, the majority of jobs require at least some post-secondary education. This includes manufacturing jobs that in the past would have afforded workers a middle class income with only a high school diploma. Those jobs are increasingly scarce. This phenomenon has also no doubt contributed to the increasing inequality of income that we observe in the U.S. today.
Why analyze the distribution of income?

In a market economy, labor markets work efficiently to match job seekers with employers needing their skills who, in turn, pay wages and salaries based on the value that workers bring to firms. Since not everyone has the same job skills, labor markets result in considerable income inequality.

In 2016, the median American household income was $59,039 (the median is the level where half of all families had more than that level and half had less). At the same time, the lowest quintile of American households (a quintile is one fifth or 20%) earned no more $24,002, while the highest quintile earned at least $121,019. While income inequality can motivate people to work harder and improve their skills, recent evidence suggests that if income inequality gets too extreme, it can adversely affect the functioning of the economy as a whole.

Labor markets do not take into account how much income a family needs for food, shelter, clothing, and health care. Market forces do not worry about what happens to families when a major local employer goes out of business. Market forces do not take time to contemplate whether those who are earning higher incomes should pay an even higher share of taxes.

According to the U.S. Census Bureau, in 2016 the federal government classified almost 41 million Americans as living with family incomes below the poverty line. Think about a family of three—perhaps a single mother with two children—attempting to pay for the basics of life on perhaps $17,916 per year. After paying for rent, healthcare, clothing, and transportation, such a family might have $6,000 to spend on food. Spread over 365 days, the food budget for the entire family would be about $17 per day. To put this in perspective, most cities have restaurants where $17 will buy you an appetizer for one.

This module begins by exploring how the U.S. government defines poverty, the balance between assisting the poor without discouraging work, and how federal antipoverty programs work. The module also discusses income inequality—
how economists measure inequality, why inequality has changed in recent decades, the range of possible government policies to reduce inequality, and the danger of a tradeoff that too great a reduction in inequality may reduce incentives for producing output.

In this module, we will consider questions such as:

- What can be done to reduce the amount of poverty?
- How does the social safety net result in a poverty trap?
- Can we, and should we, reduce the amount of economic inequality?
- If so, what’s the best way?

INTRODUCTION TO POVERTY

What you’ll learn to do: explain poverty and the poverty trap

Firms hire workers because they value the workers’ productivity. Employees get paid based on the value of their productivity to their employer. Labor productivity depends on a worker’s talents, skills and abilities. This means that people who lack marketable skills tend to be qualified only for low wage jobs, if they are employed at all.

Poverty is what we call the condition of people who do not earn enough income to be able to afford the necessities of life, which is measured by the poverty line. The poverty rate is what percentage of the population lives below the poverty line. In this section, we will examine poverty. Later we will take a closer look at income inequality, which refers to the disparity between those with higher and lower incomes.
Introduction to Poverty

The labor markets that determine what workers are paid do not take into account how much income a family needs for food, shelter, clothing, and health care. Market forces do not worry about what happens to families when a major local employer goes out of business. Market forces do not take time to contemplate whether those who are earning higher incomes should pay an even higher share of taxes.

However, labor markets do create considerable income inequalities. In 2014, the median American family income was $57,939 (the median is the level where half of all families had more than that level and half had less). According to the U.S. Census Bureau, the federal government classified almost nine million U.S. families as below the poverty line in that year. Think about a family of three—perhaps a single mother with two children—attempting to pay for the basics of life on perhaps $17,916 per year. After paying for rent, healthcare, clothing, and transportation, such a family might have $6,000 to spend on food. Spread over 365 days, the food budget for the entire family would be about $17 per day. To put this in perspective, most cities have restaurants where $17 will buy you an appetizer for one.

This section explores how the U.S. government defines poverty, the balance between assisting the poor without discouraging work, and how federal antipoverty programs work.

Drawing the Poverty Line

Comparisons of high and low incomes raise two different issues: economic inequality and poverty. Poverty is measured by the number of people who fall below a certain level of income—called the poverty line—that defines the income needed for a basic standard of living. Income inequality compares the share of the total income (or wealth) in society that is received by different groups; for example, comparing the share of income received by the top 10% to the share of income received by the bottom 10%.

In the United States, the official definition of the poverty line traces back to a single person: Mollie Orshansky. In 1963, Orshansky, who was working for the Social Security Administration, published an article called “Children of the Poor” in a highly useful and dry-as-dust publication called the Social Security Bulletin. Orshansky’s idea was to define a poverty line based on the cost of a healthy diet.

Her previous job had been at the U.S. Department of Agriculture, where she had worked in an agency called the Bureau of Home Economics and Human Nutrition. One task of this bureau had been to calculate how much it would cost to feed a nutritionally adequate diet to a family. Orshansky found that the average family spent one-third of its income on food. She then proposed that the poverty line be the amount needed to buy a nutritionally adequate diet, given the size of the family, multiplied by three.

The current U.S. poverty line is essentially the same as the Orshansky poverty line, although the government adjusts the dollar amounts to represent the same buying power over time. The U.S. poverty line in 2015 ranged from $11,790 for a single individual to $25,240 for a household of four people.

Figure 1 shows the U.S. poverty rate over time; that is, the percentage of the population below the poverty line in any given year. The poverty rate declined through the 1960s, rose in the early 1980s and early 1990s, but seems to have
been slightly lower since the mid-1990s. However, in no year in the last four decades has the poverty rate been less than 11% of the U.S. population—that is, at best about one American in nine is below the poverty line. In recent years, the poverty rate appears to have peaked at 15.9% in 2011 before dropping to 14.5% in 2013.

Figure 1. The U.S. Poverty Rate since 1960. The poverty rate fell dramatically during the 1960s, rose in the early 1980s and early 1990s, and, after declining in the 1990s through mid-2000s, rose to 15.9% in 2011, which is close to the 1960 levels. In 2013, the poverty dropped slightly to 14.5%.
(Source: U.S. Census Bureau)

Table 1 compares poverty rates for different groups in 2013. As you will see when we delve further into these numbers, poverty rates are relatively low for whites, for the elderly, for the well-educated, and for male-headed households. Poverty rates for females, Hispanics, and African Americans are much higher than for whites. While Hispanics and African Americans have a higher percentage of individuals living in poverty than others, most people in the United States living below the poverty line are white.

<table>
<thead>
<tr>
<th>Group</th>
<th>Poverty Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Females</td>
<td>15.8%</td>
</tr>
<tr>
<td>Males</td>
<td>13.1%</td>
</tr>
<tr>
<td>White</td>
<td>9.6%</td>
</tr>
<tr>
<td>Black</td>
<td>27.1%</td>
</tr>
<tr>
<td>Hispanic</td>
<td>23.5%</td>
</tr>
<tr>
<td>Under age 18</td>
<td>19.9%</td>
</tr>
<tr>
<td>Ages 18–24</td>
<td>20.6%</td>
</tr>
<tr>
<td>Ages 25–34</td>
<td>15.9%</td>
</tr>
<tr>
<td>Ages 35–44</td>
<td>12.2%</td>
</tr>
<tr>
<td>Ages 45–54</td>
<td>10.9%</td>
</tr>
<tr>
<td>Ages 55–59</td>
<td>10.7%</td>
</tr>
<tr>
<td>Ages 60–64</td>
<td>10.8%</td>
</tr>
<tr>
<td>Ages 65 and older</td>
<td>9.5%</td>
</tr>
</tbody>
</table>
The concept of a poverty line raises many tricky questions. In a vast country like the United States, should there be a national poverty line? After all, according to the Federal Register, the median household income for a family of four was $102,552 in New Jersey and $57,132 in Mississippi in 2013, and prices of some basic goods like housing are quite different between states. The poverty line is based on cash income, which means it does not account for government programs that provide assistance to the poor in a non-cash form, like Medicaid (health care for low-income individuals and families) and food aid. Also, low-income families can qualify for federal housing assistance. (We will discuss these and other government aid programs in detail later in this chapter.)

Should the government adjust the poverty line to account for the value of such programs? Many economists and policymakers wonder whether we should rethink the concept of what poverty means in the twenty-first century.

**HOW IS POVERTY MEASURED IN LOW-INCOME COUNTRIES?**

The World Bank sets two poverty lines for low-income countries around the world. One poverty line is set at an income of $1.25/day per person; the other is at $2/day. By comparison, the U.S. 2011 poverty line of $17,916 annually for a family of three works out to $16.37 per person per day.

Clearly, many people around the world are far poorer than Americans, as Table 14.2 shows. China and India both have more than a billion people; Nigeria is the most populous country in Africa; and Egypt is the most populous country in the Middle East. In all four of those countries, in the mid-2000s, a substantial share of the population subsisted on less than $2/day. Indeed, about half the world lives on less than $2.50 a day, and 80 percent of the world lives on less than $10 per day. (Of course, the cost of food, clothing, and shelter in those countries can be very different from those costs in the United States, so the $2 and $2.50 figures may mean greater purchasing power than they would in the United States.)

<table>
<thead>
<tr>
<th>Country (in year)</th>
<th>Share of Population below $1.25/Day</th>
<th>Share of Population below $2.00/Day</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brazil (in 2009)</td>
<td>6.1%</td>
<td>10.8%</td>
</tr>
<tr>
<td>China (in 2009)</td>
<td>11.8%</td>
<td>27.2%</td>
</tr>
<tr>
<td>Egypt (in 2008)</td>
<td>1.7%</td>
<td>15.4%</td>
</tr>
<tr>
<td>India (in 2010)</td>
<td>32.7%</td>
<td>68.8%</td>
</tr>
<tr>
<td>Mexico (in 2010)</td>
<td>0.7%</td>
<td>4.5%</td>
</tr>
<tr>
<td>Nigeria (in 2010)</td>
<td>68.0%</td>
<td>84.5%</td>
</tr>
</tbody>
</table>

Any poverty line will be somewhat arbitrary, and it is useful to have a poverty line whose basic definition does not change much over time. If Congress voted every few years to redefine what poverty means, then it would be difficult to compare rates over time. After all, would a lower poverty rate mean that the definition had been changed, or that people were actually better off? Government statisticians at the U.S. Census Bureau have ongoing research programs to address questions like these.

**Try It**

Visit this page in your course online to check your understanding.

**Glossary**

income inequality: when one group receives a disproportionate share of total income or wealth than others

poverty: the situation of being below a certain level of income one needs for a basic standard of living
poverty line: the specific amount of income one requires for a basic standard of living

poverty rate: percentage of the population living below the poverty line

THE POVERTY TRAP

Learning Objectives

• Explain the poverty trap, noting how government programs impact it

The Poverty Trap

Can you give people too much help, or the wrong kind of help? When people are provided with food, shelter, healthcare, income, and other necessities, assistance may reduce their incentive to work. Consider a program to fight poverty that works in this reasonable-sounding manner: the government provides assistance to the poor, but as the poor earn income to support themselves, the government reduces the level of assistance it provides. With such a program, every time a poor person earns $100, the person loses $100 in government support. As a result, the person experiences no net gain for working. Economists call this problem the poverty trap.

Consider the situation a single-parent family faces. Figure 1 illustrates a single mother (earning $8 an hour) with two children. First, consider the labor-leisure budget constraint that this family faces in a situation without government assistance. On the horizontal axis is hours of leisure (or time spent with family responsibilities) increasing in quantity from right to left. Also on the horizontal axis is the number of hours at paid work, going from zero hours on the right to the maximum of 2,500 hours on the left. On the vertical axis is the amount of income per year rising from low to higher amounts of income. The budget constraint line shows that at zero hours of leisure and 2,500 hours of work, the maximum amount of income is $20,000 ($8 × 2,500 hours). At the other extreme of the budget constraint line, an individual would work zero hours, earn zero income, but enjoy 2,500 hours of leisure. At point A on the budget constraint line, by working 40 hours a week, 50 weeks a year, the utility-maximizing choice is to work a total of 2,000 hours per year and earn $16,000.
Figure 1. The Poverty Trap in Action. The original choice is 500 hours of leisure, 2,000 hours of work at point A, and income of $16,000. With a guaranteed income of $18,000, this family would receive $18,000 whether it provides zero hours of work or 2,000 hours of work. Only if the family provides, say, 2,300 hours of work does its income rise above the guaranteed level of $18,000—and even then, the marginal gain to income from working many hours is small.

Now suppose that a government antipoverty program guarantees every family with a single mother and two children $18,000 in income. This is represented on the graph by a horizontal line at $18,000. With this program, each time the mother earns $1,000, the government will deduct $1,000 of its support. Table 1 shows what will happen at each combination of work and government support.

Table 1. Total Income at Various Combinations of Work and Support

<table>
<thead>
<tr>
<th>Amount Worked (hours)</th>
<th>Total Earnings</th>
<th>Government Support</th>
<th>Total Income</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>$18,000</td>
<td>$18,000</td>
</tr>
<tr>
<td>500</td>
<td>$4,000</td>
<td>$14,000</td>
<td>$18,000</td>
</tr>
<tr>
<td>1,000</td>
<td>$8,000</td>
<td>$10,000</td>
<td>$18,000</td>
</tr>
<tr>
<td>1,500</td>
<td>$12,000</td>
<td>$6,000</td>
<td>$18,000</td>
</tr>
<tr>
<td>2,000</td>
<td>$16,000</td>
<td>$2,000</td>
<td>$18,000</td>
</tr>
<tr>
<td>2,500</td>
<td>$20,000</td>
<td>0</td>
<td>$20,000</td>
</tr>
</tbody>
</table>

The new budget line, with the antipoverty program in place, is the horizontal and heavy line that is flat at $18,000. If the mother does not work at all, she receives $18,000, all from the government. If she works full time, giving up 40 hours per week with her children, she still ends up with $18,000 at the end of the year. Only if she works 2,300 hours in the year—which is an average of 44 hours per week for 50 weeks a year—does household income rise to $18,400. Even in this case, all of her year’s work means that household income rises by only $400 over the income she would receive if she did not work at all. She would need to work 50 hours a week to reach $20,000.

Indeed, the poverty trap is even stronger than this simplified example shows, because a working mother will have extra expenses like clothing, transportation, and child care that a nonworking mother will not face, making the economic gains from working even smaller. Moreover, those who do not work fail to build up job experience and contacts, which makes working in the future even less likely.

The bite of the poverty trap can be reduced by designing an antipoverty program so that, instead of reducing government payments by $1 for every $1 earned, payments are reduced by some smaller amount instead. The bite of
the poverty trap can also be reduced by imposing requirements for work as a condition of receiving benefits and setting a time limit on benefits.

Figure 2 illustrates a government program that guarantees $18,000 in income, even for those who do not work at all, but then reduces this amount by 50 cents for each $1 earned. The new, higher budget line in Figure 2 shows that, with this program, additional hours of work will bring some economic gain. Because of the reduction in government income when an individual works, an individual earning $8.00 will really net only $4.00 per hour. The vertical intercept of this higher budget constraint line is at $28,000 ($18,000 + 2,500 hours × $4.00 = $28,000). The horizontal intercept is at the point on the graph where $18,000 and 2500 hours of leisure is set.

Table 2 shows the total income differences with various choices of labor and leisure.

<table>
<thead>
<tr>
<th>Amount Worked (hours)</th>
<th>Total Earnings</th>
<th>Government Support</th>
<th>Total Income</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>$18,000</td>
<td>$18,000</td>
</tr>
<tr>
<td>500</td>
<td>$4,000</td>
<td>$16,000</td>
<td>$20,000</td>
</tr>
<tr>
<td>1,000</td>
<td>$8,000</td>
<td>$14,000</td>
<td>$22,000</td>
</tr>
<tr>
<td>1,500</td>
<td>$12,000</td>
<td>$12,000</td>
<td>$24,000</td>
</tr>
<tr>
<td>2,000</td>
<td>$16,000</td>
<td>$10,000</td>
<td>$26,000</td>
</tr>
<tr>
<td>2,500</td>
<td>$20,000</td>
<td>$8,000</td>
<td>$28,000</td>
</tr>
</tbody>
</table>
However, this type of program raises other issues. First, even if it does not eliminate the incentive to work by reducing government payments by $1 for every $1 earned, enacting such a program may still reduce the incentive to work. At least some people who would be working 2,000 hours each year without this program might decide to work fewer hours but still end up with more income—that is, their choice on the new budget line would be like S, above and to the right of the original choice P. Of course, others may choose a point like R, which involves the same amount of work as P, or even a point to the left of R that involves more work.

The second major issue is that when the government phases out its support payments more slowly, the antipoverty program costs more money. Still, it may be preferable in the long run to spend more money on a program that retains a greater incentive to work, rather than spending less money on a program that nearly eliminates any gains from working.

There are a variety of government support programs focused specifically on the poor, including TANF (the basic welfare program), SNAP (food stamps), Medicaid, and the earned income tax credit (EITC). These programs will be described in more detail in the next section. Although the programs vary from state to state, it is generally a true statement that in many states from the 1960s into the 1980s, if poor people worked, their level of income barely rose—or did not rise at all—after the reduction in government support payments was factored in. The following feature shows how this happens.

**CALCULATING A BUDGET CONSTRAINT LINE**

Jason earns $9.00 an hour, and a government antipoverty program provides a floor of $10,000 guaranteed income. The government reduces government support by $0.50 for each $1.00 earned. What are the horizontal and vertical intercepts of the budget constraint line? Assume the maximum hours for work or leisure is 2,500 hours.

**Step 1.** Determine the amount of the government guaranteed income. In this case, it is $10,000.

**Step 2.** Plot that guaranteed income as a horizontal line on the budget constraint line.

**Step 3.** Determine what Jason earns if he has no income and enjoys 2,500 hours of leisure. In this case, he will receive the guaranteed $10,000 (the horizontal intercept).

**Step 4.** Calculate how much Jason’s salary will be reduced by due to the reduction in government income. In Jason’s case, it will be reduced by one half. He will, in effect, net only $4.50 an hour.

**Step 5.** If Jason works 1,000 hours, at a maximum what income will Jason receive? Jason will get the government assistance of $10,000. He will net only $4.50 for every hour he chooses to work. If he works 1,000 hours at $4.50, his earned income is $4,500 plus the government income of $10,000. Thus the total maximum income (the vertical intercept) is $10,000 + $4,500 = $14,500.

**Try It**

Visit this page in your course online to check your understanding.

glossary

**poverty trap:** antipoverty programs set up so that government benefits decline substantially as people earn more income—as a result, working provides little financial gain

---

**LEARN BY DOING: THE POVERTY TRAP**
The Safety Net

The U.S. government has implemented a number of programs to assist those below the poverty line and those who have incomes just above the poverty line, who are referred to as the near-poor. Such programs are called the safety net, in recognition of the fact that they offer some protection for those who find themselves without jobs or income.

Temporary Assistance for Needy Families

From the Great Depression of the 1930s until 1996, the United States’ most visible antipoverty program was Aid to Families with Dependent Children (AFDC), which provided cash payments to mothers with children whose income was below the poverty line. This program was often just called “welfare.” In 1996, Congress passed and President Bill Clinton signed into law the Personal Responsibility and Work Opportunity Reconciliation Act, more commonly called the welfare reform act. The new law replaced AFDC with Temporary Assistance for Needy Families (TANF).

TANF brought several dramatic changes in how welfare operated. Under the old AFDC program, states set the level of welfare benefits that they would pay to the poor, and the federal government guaranteed it would chip in some of the money as well. The federal government’s welfare spending would rise or fall depending on the number of poor people, and on how each state set its own welfare contribution.

Under TANF, however, the federal government gives a fixed amount of money to each state. The state can then use the money for almost any program with an antipoverty component: for example, the state might use the money to give cash to poor families, or to reduce teenage pregnancy, or even to raise the high school graduation rate. However, the federal government imposed two key requirements. First, if states are to keep receiving the TANF grants, they must impose work requirements so that most of those receiving TANF benefits are working (or attending school). Second, no one can receive TANF benefits with federal money for more than a total of five years over his or her lifetime. The old AFDC program had no such work requirements or time limits.
TANF attempts to avoid the poverty trap by requiring that welfare recipients work and by limiting the length of time they can receive benefits. In its first few years, the program was quite successful. The number of families receiving payments in 1995, the last year of AFDC, was 4.8 million. By 2012, according to the Congressional Research Service, the average number of families receiving payments under TANF was 1.8 million—a decline of more than half.

TANF benefits to poor families vary considerably across states. For example, again according to the Congressional Research Service, in 2011 the highest monthly payment in Alaska to a single mother with two children was $923, while in Mississippi the highest monthly payment to that family was $170. These payments reflect differences in states’ cost of living. Total spending on TANF was approximately $16.6 billion in 1997. As of 2012, spending was at $12 billion, an almost 28% decrease, split about evenly between the federal and state governments. When you take into account the effects of inflation, the decline is even greater. Moreover, there seemed little evidence that poor families were suffering a reduced standard of living as a result of TANF—although, on the other side, there was not much evidence that poor families had greatly improved their total levels of income, either.

The Earned Income Tax Credit (EITC)

The earned income tax credit (EITC), first passed in 1975, is a method of assisting the working poor through the tax system. The EITC is the second largest assistance program for low-income groups (after SNAP, described below), and projections for 2013 expected 26 million households to take advantage of it at an estimated cost of $50 billion. In 2013, for example, a single parent with two children would have received a tax credit of $5,372 up to an income level of $17,530. The amount of the tax break increases with the amount of income earned, up to a point. The earned income tax credit has often been popular with both economists and the general public because of the way it effectively increases the payment received for work.

What about the danger of the poverty trap that every additional $1 earned will reduce government support payments by close to $1? To minimize this problem, the earned income tax credit is phased out slowly. According to the Tax Policy Center, for a single-parent family with two children in 2013, the credit is not reduced at all (but neither is it increased) as earnings rise from $13,430 to $17,530. Then, for every $1 earned above $17,530, the amount received from the credit is reduced by 21.06 cents, until the credit phases out completely at an income level of $46,227.

Figure 1 illustrates that the earned income tax credits, child tax credits, and the TANF program all cost the federal government money—either in direct outlays or in loss of tax revenues. CTC stands for the government child tax credit.

Figure 1. Real Federal Spending on CTC, EITC, and TANF, 1975-2013. EITC increased from more than $20 billion in 2000 to over an estimated $50 billion by 2013, far exceeding estimated 2013 outlays in the CTC (Child Tax Credits) and TANF of over $20 billion and $10 billion, respectively. (Source: Office of Management and Budget)

In recent years, the EITC has become the single most expensive government program for providing income assistance to the poor and near-poor, costing about $60 billion in 2012. In that year, the EITC provided benefits to about 27 million families and individuals and, on average, is worth about $2,296 per family (with children), according to the Tax Policy
Center. One reason that the TANF law worked as well as it did is that the EITC was greatly expanded in the late 1980s and again in the early 1990s, which increased the returns to work for low-income Americans.

Supplemental Nutrition Assistance Program (SNAP)

Often called “food stamps,” Supplemental Nutrition Assistance Program (SNAP) is a federally funded program, started in 1964, in which each month poor people receive a card like a debit card that they can use to buy food. The amount of food aid for which a household is eligible varies by income, number of children, and other factors but, in general, households are expected to spend about 30% of their own net income on food, and if 30% of their net income is not enough to purchase a nutritionally adequate diet, then those households are eligible for SNAP.

SNAP can contribute to the poverty trap. For every $100 earned, the government assumes that a family can spend $30 more for food, and thus reduces its eligibility for food aid by $30. This decreased benefit is not a complete disincentive to work—but combined with how other programs reduce benefits as income increases, it adds to the problem. SNAP, however, does try to address the poverty trap with its own set of work requirements and time limits.

Why give debit cards and not just cash? Part of the political support for SNAP comes from a belief that since the cards must be spent on food, they cannot be “wasted” on other forms of consumption. From an economic point of view, however, the belief that cards must increase spending on food seems wrong-headed. After all, say that a poor family is spending $2,500 per year on food, and then it starts receiving $1,000 per year in SNAP aid. The family might react by spending $3,500 per year on food (income plus aid), or it might react by continuing to spend $2,500 per year on food, but use the $1,000 in food aid to free up $1,000 that can now be spent on other goods. So it is reasonable to think of SNAP cards as an alternative method, along with TANF and the earned income tax credit, of transferring income to the working poor.

Indeed, anyone eligible for TANF is also eligible for SNAP, although states can expand eligibility for food aid if they wish to do so. In some states, where TANF welfare spending is relatively low, a poor family may receive more in support from SNAP than from TANF. In 2012, about 46.6 million people received food aid at an annual cost of about $74.6 billion, with an average monthly benefit of about $287 per person per month. SNAP participation increased by 70% between 2007 and 2011, from 26.6 million participants to 45 million. According to the Congressional Budget Office, this dramatic rise in participation was caused by the Great Recession of 2008–2009 and rising food prices.

The federal government deploys a range of income security programs that are funded through departments such as Health and Human Services, Agriculture, and Housing and Urban Development (HUD) (see Figure 14.6). According to the Office of Management and Budget, collectively, these three departments provided an estimated $62 billion of aid through programs such as supplemental feeding programs for women and children, subsidized housing, and energy assistance. The federal government also transfers funds to individual states through special grant programs.
The safety net includes a number of other programs: government-subsidized school lunches and breakfasts for children from low-income families; the Special Supplemental Food Program for Women, Infants and Children (WIC), which provides food assistance for pregnant women and newborns; the Low Income Home Energy Assistance Program, which provides help with home heating bills; housing assistance, which helps pay the rent; and Supplemental Security Income, which provides cash support for the disabled and the elderly poor.

**Medicaid**

Medicaid was created by Congress in 1965 and is a joint health insurance program entered into by both the states and the federal government. The federal government helps fund Medicaid, but each state is responsible for administering the program, determining the level of benefits, and determining eligibility. It provides medical insurance for certain low-income people, including those below the poverty line, with a focus on families with children, the elderly, and the disabled. About one-third of Medicaid spending is for low-income mothers with children. An increasing share of the program funding in recent years has gone to pay for nursing home costs for the elderly poor. The program ensures that a basic level of benefits is provided to Medicaid participants, but because each state sets eligibility requirements and provides varying levels of service, the program differs from state to state.

In the past, a common problem has been that many low-paying jobs pay enough to a breadwinner so that a family could lose its eligibility for Medicaid, yet the job does not offer health insurance benefits. A poor parent considering such a job might choose not to work rather than lose health insurance for his or her children. In this way, health insurance can become a part of the poverty trap. Many states recognized this problem in the 1980s and 1990s and expanded their Medicaid coverage to include not just the poor, but the near-poor earning up to 135% or even 185% of the poverty line. Some states also guaranteed that children would not lose coverage if their parents worked.

These expanded guarantees cost the government money, of course, but they also helped to encourage those on welfare to enter the labor force. As of 2012, approximately 67 million people participated in Medicaid. Of those enrolled, almost half are children. Healthcare expenditures, however, are highest for the elderly population, which comprises approximately 25% of participants. As Figure 3(a) indicates, the largest number of households that enroll in Medicaid are those with children. Lower-income adults are the next largest group enrolled in Medicaid at 28%. The blind and disabled are 16% of those enrolled, and seniors are 9% of those enrolled. Figure 3(b) shows how much actual Medicaid dollars are spent for each group. Out of total Medicaid spending, more is spent on seniors (20%) and the blind and
disabled (44%). So, 64% of all Medicaid spending goes to seniors, the blind, and disabled. Children receive 21% of all Medicaid spending, followed by adults at 15%.

Figure 3. Medicaid Enrollment and Spending. Part (a) shows the Medicaid enrollment by different populations, with children comprising the largest percentage at 47%, followed by adults at 28%, and the blind and disabled at 16%. Part (b) shows that Medicaid spending is principally for the blind and disabled, followed by the elderly. Although children are the largest population covered by Medicaid, expenditures on children are only at 21%.

Try It
Visit this page in your course online to check your understanding.

Glossary

**earned income tax credit (EITC):** a method of assisting the working poor through the tax system

**Medicaid:** a federal–state joint program enacted in 1965 that provides medical insurance for certain (not all) low-income people, including the near-poor as well as those below the poverty line, and focusing on low-income families with children, the low-income elderly, and the disabled

**near-poor:** those who have incomes just above the poverty line

**safety net:** the group of government programs that provide assistance to the poor and the near-poor

**Supplemental Nutrition Assistance Program (SNAP):** a federally funded program, started in 1964, in which each month poor people receive SNAP cards they can use to buy food

**Temporary Assistance for Needy Families (TANF):** government assistance program in which the federal government gives money to states to help provide assistance to those in poverty. People receiving assistance are required to find a job within two years and cannot receive benefits for more than five years.

Licensing & Attributions

CC licensed content, Shared previously

- The Safety Net. **Authored by:** OpenStax College. Located at: https://cnx.org/contents/vEm0H...p@84.41:c7Ll0Z_9@2/The-Safety-Net. **License:** CC BY Attribution. **License Terms:** Download for free at http://cnx.org/content/col11627/latest
In September 2011, a group of protesters gathered in Zuccotti Park in New York City to decry what they perceived as increasing social and economic inequality in the United States. Calling their protest “Occupy Wall Street,” they argued that the concentration of wealth among the richest 1% in the United States was both economically unsustainable and inequitable, and needed to be changed. The protest then spread to other major cities, and the Occupy movement was born.

Why were people so upset? How much wealth is concentrated among the top 1% in our society? How did they acquire so much wealth? These are very real, very important questions in the United States now, and this section on economic inequality will help us address the causes behind this sentiment.
In a market economy, your income depends on the resources you own (e.g. labor, land, etc.), and the value the market places on those resources. People who own a lot of resources and people who own resources that are highly valued will tend to earn higher incomes than people who do not. As a consequence, market economies tend to result in inequality of income and wealth. Whether this is good or bad depends at least in part on the degree of inequality. Few Americans believe that Bill Gates doesn't deserve to be rich, because of the significant value his company, Microsoft, has brought to people. But should he have 100 times the wealth of the average American or 1 million times? That is the question.

Poverty levels can be subjective based on the overall income levels of a country. Typically a government measures poverty based on a percentage of the median income. Income inequality, however, has to do with the distribution of that income, in terms of which group receives the most or the least income. Income inequality involves comparing those with high incomes, middle incomes, and low incomes—not just looking at those below or near the poverty line. In turn, measuring income inequality means dividing the population into various groups and then comparing the groups, a task that we can carry out in several ways.
Poverty can change even when inequality does not move at all. Imagine a situation in which income for everyone in the population declines by 10%. Poverty would rise, since a greater share of the population would now fall below the poverty line. However, inequality would be the same, because everyone suffered the same proportional loss. Conversely, a general rise in income levels over time would keep inequality the same, but reduce poverty.

It is also possible for income inequality to change without affecting the poverty rate. Imagine a situation in which a large number of people who already have high incomes increase their incomes by even more. Inequality would rise as a result—but the number of people below the poverty line would remain unchanged.

Why did inequality of household income increase in the United States in recent decades? Indeed, a trend toward greater income inequality has occurred in many countries around the world, although the effect has been more powerful in the U.S. economy. Economists have focused their explanations for the increasing inequality on two factors that changed more or less continually from the 1970s into the 2000s. One set of explanations focuses on the changing shape of American households; the other focuses on greater inequality of wages, what some economists call “winner take all” labor markets. We will begin with how we measure inequality, and then consider the explanations for growing inequality in the United States.

**Measuring Income Distribution by Quintiles**

One common way of measuring income inequality is to rank all households by income, from lowest to highest, and then to divide all households into five groups with equal numbers of people, known as quintiles. This calculation allows for measuring the distribution of income among the five groups compared to the total. The first quintile is the lowest fifth or
20%, the second quintile is the next lowest, and so on. We can measure income inequality by comparing what share of the total income each quintile earns.

U.S. income distribution by quintile appears in Table 1. In 2016, for example, the bottom quintile of the income distribution received 3.1% of income; the second quintile received 8.3%; the third quintile, 14.2%; the fourth quintile, 22.9%; and the top quintile, 51.5%. The final column of Figure 1 shows what share of income went to households in the top 5% of the income distribution: 22.6% in 2016. Over time, from the late 1960s to the early 1980s, the top fifth of the income distribution typically received between about 43% to 44% of all income. The share of income that the top fifth received then begins to rise. Census Bureau researchers trace, much of this increase in the share of income going to the top fifth to an increase in the share of income going to the top 5%. The quintile measure shows how income inequality has increased in recent decades.

Table 1. Share of Aggregate Income Received by Each Fifth and Top 5 Percent of Households, All Races: 1967 to 2016 (Source: U.S. Census Bureau, Table H-2)

<table>
<thead>
<tr>
<th>Year</th>
<th>Shares of aggregate income</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Lowest fifth</td>
</tr>
<tr>
<td>1967</td>
<td>4.0</td>
</tr>
<tr>
<td>1970</td>
<td>4.1</td>
</tr>
<tr>
<td>1975</td>
<td>4.3</td>
</tr>
<tr>
<td>1980</td>
<td>4.2</td>
</tr>
<tr>
<td>1985</td>
<td>3.9</td>
</tr>
<tr>
<td>1990</td>
<td>3.8</td>
</tr>
<tr>
<td>1995</td>
<td>3.7</td>
</tr>
<tr>
<td>2000</td>
<td>3.6</td>
</tr>
<tr>
<td>2005</td>
<td>3.4</td>
</tr>
<tr>
<td>2010</td>
<td>3.3</td>
</tr>
<tr>
<td>2015</td>
<td>3.1</td>
</tr>
<tr>
<td>2016</td>
<td>3.1</td>
</tr>
</tbody>
</table>

Source: U.S. Census Bureau, Current Population Survey, Annual Social and Economic Supplements. For information on confidentiality protection, sampling error, nonsampling error, and definitions, see //www2.census.gov/programs-surveys/cps/techdocs/cpsmar17.pdf

It can also be useful to divide the income distribution in ways other than quintiles; for example, into tenths or even into percentiles (that is, hundredths). A more detailed breakdown can provide additional insights. For example, the last column of Table 1 shows the income received by the top 5% percent of the income distribution. Between 1980 and 2016, the share of income going to the top 5% increased by 6.1 percentage points (from 16.5% in 1980 to 22.6% in 2016). From 1980 to 2016 the share of income going to the top quintile increased by 7.5 percentage points (from 44.1% in 1980 to 51.5% in 2016). Thus, the top 20% of householders (the fifth quintile) received over half (51.5%) of all the income in the United States in 2016.

Try It

Visit this page in your course online to check your understanding.
Lorenz Curve

The data on income inequality can be presented in various ways. For example, you could draw a bar graph that showed the share of income going to each fifth of the income distribution. Figure 1 presents an alternative way of showing inequality data in what is called a **Lorenz curve**. The Lorenz curve shows the cumulative share of population on the horizontal axis and the cumulative percentage of total income received on the vertical axis.

![Figure 2. The Lorenz Curve](image)

Every Lorenz curve diagram begins with a line sloping up at a 45-degree angle. We show it as a dashed line in Figure 2. The points along this line show what perfect equality of the income distribution looks like. It would mean, for example, that the bottom 20% of the income distribution receives 20% of the total income, the bottom 40% gets 40% of total income, and so on. The other lines reflect actual U.S. data on inequality for 1980 and 2016.

The trick in graphing a Lorenz curve is that you must change the shares of income for each specific quintile, which we show in the first column of numbers in Table 2, into cumulative income, which we show in the second column of numbers. For example, the bottom 40% of the cumulative income distribution will be the sum of the first and second quintiles; the bottom 60% of the cumulative income distribution will be the sum of the first, second, and third quintiles, and so on. The final entry in the cumulative income column needs to be 100%, because by definition, 100% of the population receives 100% of the income.

**Table 2. Calculating the Lorenz Curve**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>First quintile</td>
<td>4.2</td>
<td>4.2</td>
<td>3.1</td>
<td>3.1</td>
</tr>
</tbody>
</table>
In a Lorenz curve diagram, a more unequal distribution of income will loop farther down and away from the 45-degree line, while a more equal distribution of income will move the line closer to the 45-degree line. Figure 2 illustrates the greater inequality of the U.S. income distribution between 1980 and 2016 because the Lorenz curve for 2016 is farther from the 45-degree line than for 1980. The Lorenz curve is a useful way of presenting the quintile data that provides an image of all the quintile data at once.

Measuring Income Inequality

The U.S. economy has a relatively high degree of income inequality by global standards. As Table 3 shows, based on a variety of national surveys done for a selection of years in the last five years of the 2000s (with the exception of Germany, and adjusted to make the measures more comparable), the U.S. economy has greater inequality than Germany (along with most Western European countries). The region of the world with the highest level of income inequality is Latin America, illustrated in the numbers for Brazil and Mexico. The level of inequality in the United States is lower than in some of the low-income countries of the world, like China and Nigeria, or some middle-income countries like the Russian Federation. However, not all poor countries have highly unequal income distributions; India provides a counterexample.


<table>
<thead>
<tr>
<th>Country</th>
<th>Survey Year</th>
<th>First Quintile</th>
<th>Second Quintile</th>
<th>Third Quintile</th>
<th>Fourth Quintile</th>
<th>Fifth Quintile</th>
</tr>
</thead>
<tbody>
<tr>
<td>United States</td>
<td>2011</td>
<td>3.2%</td>
<td>8.4%</td>
<td>14.3%</td>
<td>23.0%</td>
<td>51.1%</td>
</tr>
<tr>
<td>Germany</td>
<td>2000</td>
<td>8.5%</td>
<td>13.7%</td>
<td>17.8%</td>
<td>23.1%</td>
<td>36.9%</td>
</tr>
<tr>
<td>Brazil</td>
<td>2009</td>
<td>2.9%</td>
<td>7.1%</td>
<td>12.4%</td>
<td>19.0%</td>
<td>58.6%</td>
</tr>
<tr>
<td>Mexico</td>
<td>2010</td>
<td>4.9%</td>
<td>8.8%</td>
<td>13.3%</td>
<td>20.2%</td>
<td>52.8%</td>
</tr>
<tr>
<td>China</td>
<td>2009</td>
<td>4.7%</td>
<td>9.7%</td>
<td>15.3%</td>
<td>23.2%</td>
<td>47.1%</td>
</tr>
<tr>
<td>India</td>
<td>2010</td>
<td>8.5%</td>
<td>12.1%</td>
<td>15.7%</td>
<td>20.8%</td>
<td>42.8%</td>
</tr>
<tr>
<td>Russia</td>
<td>2009</td>
<td>6.1%</td>
<td>10.4%</td>
<td>14.8%</td>
<td>21.3%</td>
<td>47.1%</td>
</tr>
<tr>
<td>Nigeria</td>
<td>2010</td>
<td>4.4%</td>
<td>8.3%</td>
<td>13.0%</td>
<td>20.3%</td>
<td>54.0%</td>
</tr>
</tbody>
</table>

Watch It

This video explains income inequality and discusses some potential causes and fixes for reducing the large disparity between incomes in America.
CAUSES OF GROWING INEQUALITY

Learning Objectives

- Analyze the sources of income inequality and its effect on a market economy


Prior to the 1970s, most American families were supported by one wage earner. This was typically the husband, while the wife stayed home. In 1970, 41% of married women were in the labor force, but according to the Bureau of Labor Statistics, by 2015, the figure was 56.7%. We know that the average wage rate, adjusted for inflation, peaked during the
mid-1970s and has stagnated or declined since then (depending on how it is measured). This would have caused families with middle class incomes to slip down in the income distribution, if only the husband was employed. The initial impact of the increase in women's labor market participation, then, was to keep family incomes in the middle class since there were now two wage earners.

Overtime, though, a new phenomenon arose: it has become more common for one high earner to marry another high earner. A few decades ago, the common pattern featured a man with relatively high earnings, such as an executive or a doctor, marrying a woman who did not earn as much, like a secretary or a nurse. Often, the woman would leave paid employment, at least for a few years, to raise a family. However, now doctors are marrying doctors and executives are marrying executives, and mothers with high-powered careers are often returning to work while their children are quite young. This pattern of households with two high earners tends to increase the proportion of high-earning households.

According to data in the National Journal, even as two-earner couples have increased, so have single-parent households. Of all U.S. families, 13.1% were headed by single mothers. The poverty rate among single-parent households tends to be relatively high.

These changes in family structure, including the growth of single-parent families who tend to be at the lower end of the income distribution, and the growth of two-career high-earner couples near the top end of the income distribution, account for roughly half of the rise in income inequality across households in recent decades.

Watch It

When asked about the “ideal” distribution of wealth in America, majority people guessed wrong. Watch this video to see how incredibly far-off they were in their predictions and what the distribution of wealth really looks like. While the video was created several years ago, the same concerns are still applicable today.

Watch this video online: https://youtu.be/QPKKQnijnsM

Causes of Growing Inequality: A Shift in the Distribution of Wages

Another factor behind the rise in U.S. income inequality is that earnings have become less equal since the late 1970s. In particular, the earnings of high-skilled labor relative to low-skilled labor have increased. It used to be that a man could find a well paid job—one that would put his family in the middle class—with only a high school diploma. Those jobs are increasingly hard to find. By contrast, more and more jobs require at least some post-secondary education, if not a 4-year degree.

One way to measure this change is to take workers’ earnings with at least a four-year college bachelor’s degree (including those who went on and completed an advanced degree) and divide them by workers’ earnings with only a high school degree. The result is that those in the 25–34 age bracket with college degrees earned about 1.67 times as much as high school graduates in 2010, up from 1.59 times in 1995, according to U.S. Census data.

Economists use the demand and supply model to reason through the most likely causes of this shift. According to the National Center for Education Statistics, in recent decades, the supply of U.S. workers with college degrees has increased substantially. For example, 840,000 four-year bachelor’s degrees were conferred on Americans in 1970. In 2013-2014, 1,894,934 such degrees were conferred—an increase of over 90%. In Figure 1, this shift in supply to the right, from S₀ to S₁, should result in a lower equilibrium wage for high-skilled labor. Thus, we can explain the increase in the price of high-skilled labor by a greater demand, like the movement from D₀ to D₁. Evidently, combining both the increase in supply and in demand has resulted in a shift from E₀ to E₁, and a resulting higher wage.
Figure 1. Why Would Wages Rise for High-Skilled Labor? The proportion of workers attending college has increased in recent decades, so the supply curve for high-skilled labor has shifted to the right, from $S_0$ to $S_1$. If the demand for high-skilled labor had remained at $D_0$, then this shift in supply would have led to lower wages for high-skilled labor. However, the wages for high-skilled labor, especially if there is a large global demand, have increased even with the shift in supply to the right. The explanation must lie in a shift to the right in demand for high-skilled labor, from $D_0$ to $D_1$. The figure shows how a combination of the shift in supply, from $S_0$ to $S_1$, and the shift in demand, from $D_0$ to $D_1$, led to both an increase in the quantity of high-skilled labor hired and also to a rise in the wage for such labor, from $W_0$ to $W_1$.

What factors would cause the demand for high-skilled labor to rise? The most plausible explanation is that while the explosion in new information and communications technologies over the last several decades has helped many workers to become more productive, the benefits have been especially great for high-skilled workers like top business managers, consultants, and design professionals. The new technologies have also helped to encourage globalization, the remarkable increase in international trade over the last few decades, by making it more possible to learn about and coordinate economic interactions all around the world. In turn, the rising impact of foreign trade in the U.S. economy has opened up greater opportunities for high-skilled workers to sell their services around the world, and lower-skilled workers have to compete with a larger supply of similarly skilled workers around the globe.

We can view the market for high-skilled labor as a race between forces of supply and demand. Additional education and on-the-job training will tend to increase the high-skilled labor supply and to hold down its relative wage. Conversely, new technology and other economic trends like globalization tend to increase the demand for high-skilled labor and push up its relative wage. We can view the greater inequality of wages as a sign that demand for skilled labor is increasing faster than supply. Alternatively, if the supply of lower skilled workers exceeds the demand, then average wages in the lower quintiles of the income distribution will decrease. The combination of forces in the high-skilled and low-skilled labor markets leads to increased income disparity.

Try It
Visit this page in your course online to check your understanding.

Watch It
The following video explains part of the reason for the increase in income inequality in the U.S. Watch this video online: https://youtu.be/r034Rn2Sy6s
GOVERNMENT POLICIES TO REDUCE INCOME INEQUALITY

Learning Objectives

- Explain the arguments for and against government intervention to reduce economic inequality

Government Policies to Reduce Income Inequality

No society should expect or desire complete equality of income at a given point in time, for a number of reasons. First, most workers receive relatively low earnings in their first few jobs, higher earnings as they reach middle age, and then lower earnings after retirement. Thus, a society with people of varying ages will have a certain amount of income inequality. Second, people’s preferences and desires differ. Some are willing to work long hours to have income for large houses, fast cars and computers, luxury vacations, and the ability to support children and grandchildren.

These factors all imply that a snapshot of inequality in a given year does not provide an accurate picture of how people’s incomes rise and fall over time. Even if some degree of economic inequality is expected at any point in time, how much inequality should there be? There is also the difference between income and wealth, as shown in the following paragraphs.

HOW DO YOU MEASURE WEALTH VERSUS INCOME INEQUALITY?

Income is a flow of money received, often measured on a monthly or an annual basis; wealth is the sum of the value of all assets, including money in bank accounts, financial investments, a pension fund, and the value of a home. In calculating wealth all debts must be subtracted, such as debt owed on a home mortgage and on credit cards. A retired person, for example, may have relatively little income in a given year, other than a pension or Social Security. However, if that person has saved and invested over time, the person’s accumulated wealth can be quite substantial. The wealth distribution is more unequal than the income distribution, because differences in income can accumulate over time to make even larger differences in wealth. However, the degree of inequality in the wealth distribution can be measured with the same tools we use to measure the inequality in the income distribution, like quintile measurements. Data on wealth are collected once every three years in the Survey of Consumer Finance.

Even if they cannot answer the question of how much inequality is too much, economists can still play an important role in spelling out policy options and tradeoffs. If a society decides to reduce the level of economic inequality, it has three main sets of tools: redistribution from those with high incomes to those with low incomes; trying to assure that a ladder of opportunity is widely available; and a tax on inheritance.

Redistribution
Redistribution means taking income from those with higher incomes and providing income to those with lower incomes. Earlier in this module, we considered some of the key government policies that provide support for the poor: the welfare program TANF, the earned income tax credit, SNAP, and Medicaid. If a reduction in inequality is desired, these programs could receive additional funding.

The programs are paid for through the federal income tax, which is a progressive tax system designed in such a way that the rich pay a higher percent in income taxes than the poor. Data from household income tax returns in 2009 shows that the top 1% of households had an average income of $1,219,700 per year in pre-tax income and paid an average federal tax rate of 28.9%. The effective income tax, which is total taxes paid divided by total income (all sources of income such as wages, profits, interest, rental income, and government transfers such as veterans’ benefits), was much lower. The effective tax paid by the top 1% of householders was 20.4%, while the bottom two quintiles actually paid negative effective income taxes, because of provisions like the earned income tax credit. News stories occasionally report on a high-income person who has managed to pay very little in taxes, but while such individual cases exist, according to the Congressional Budget Office, the typical pattern is that people with higher incomes pay a higher average share of their income in federal income taxes.

Of course, the fact that some degree of redistribution occurs now through the federal income tax and government antipoverty programs does not settle the questions of how much redistribution is appropriate, and whether more redistribution should occur.

The Ladder of Opportunity

Economic inequality is perhaps most troubling when it is not the result of effort or talent, but instead is determined by the circumstances under which a child grows up. One child attends a well-run grade school and high school and heads on to college, while parents help out by supporting education and other interests, paying for college, a first car, and a first house, and offering work connections that lead to internships and jobs. Another child attends a poorly run grade school, barely makes it through a low-quality high school, does not go to college, and lacks family and peer support. These two children may be similar in their underlying talents and in the effort they put forth, but their economic outcomes are likely to be quite different.

Public policy can attempt to build a ladder of opportunities so that, even though all children will never come from identical families and attend identical schools, each child has a reasonable opportunity to attain an economic niche in society based on their interests, desires, talents, and efforts. Some of those initiatives include those shown in Table 1.

<table>
<thead>
<tr>
<th>Children</th>
<th>College Level</th>
<th>Adults</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Improved day care</td>
<td>• Widespread loans and grants for those in financial need</td>
<td>• Opportunities for retraining and acquiring new skills</td>
</tr>
<tr>
<td>• Enrichment programs for preschoolers</td>
<td>• Public support for a range of institutions from two-year community colleges to large research universities</td>
<td>• Prohibiting discrimination in job markets and housing on the basis of race, gender, age, and disability</td>
</tr>
<tr>
<td>• Improved public schools</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>• After school and community activities</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>• Internships and apprenticeships</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

The United States has often been called a land of opportunity. Although the general idea of a ladder of opportunity for all citizens continues to exert a powerful attraction, specifics are often quite controversial. Society can experiment with a wide variety of proposals for building a ladder of opportunity, especially for those who otherwise seem likely to start their lives in a disadvantaged position. Such policy experiments need to be carried out in a spirit of open-mindedness,
because some will succeed while others will not show positive results or will cost too much to enact on a widespread basis.

Inheritance Taxes

There is always a debate about inheritance taxes. It goes like this: on the one hand, why should people who have worked hard all their lives and saved up a substantial nest egg not be able to give their money and possessions to their children and grandchildren? In particular, it would seem un-American if children were unable to inherit a family business or a family home. On the other hand, many Americans are far more comfortable with inequality resulting from high-income people who earned their money by starting innovative new companies than they are with inequality resulting from high-income people who have inherited money from rich parents.

The United States does have an estate tax—that is, a tax imposed on the value of an inheritance—which suggests a willingness to limit how much wealth can be passed on as an inheritance. However, according to the Center on Budget and Policy Priorities, in 2013 the estate tax applied only to those leaving inheritances of more than $5.25 million and thus applies to only a tiny percentage of those with high levels of wealth.

The Tradeoff between Incentives and Income Equality

Government policies to reduce poverty or to encourage economic equality, if carried to extremes, can injure incentives for economic output. The poverty trap, for example, defines a situation where guaranteeing a certain level of income can eliminate or reduce the incentive to work. An extremely high degree of redistribution, with very high taxes on the rich, would be likely to discourage work and entrepreneurship. Thus, it is common to draw the tradeoff between economic output and equality, as shown in Figure 1(a). In this formulation, if society wishes a high level of economic output, like point A, it must also accept a high degree of inequality. Conversely, if society wants a high level of equality, like point B, it must accept a lower level of economic output because of reduced incentives for production.

This view of the tradeoff between economic output and equality may be too pessimistic, and Figure 1(b) presents an alternate vision. Here, the tradeoff between economic output and equality first slopes up, in the vicinity of choice C, suggesting that certain programs might increase both output and economic equality. For example, the policy of providing free public education has an element of redistribution, since the value of the public schooling received by children of low-income families is clearly higher than what low-income families pay in taxes. A well-educated population, however, is also an enormously powerful factor in providing the skilled workers of tomorrow and helping the economy to grow and expand. In this case, equality and economic growth may complement each other.

Moreover, policies to diminish inequality and soften the hardship of poverty may sustain political support for a market economy. After all, if society does not make some effort toward reducing inequality and poverty, the alternative might be
that people would rebel against market forces. Citizens might seek economic security by demanding that their legislators pass laws forbidding employers from ever laying off workers or reducing wages, or laws that would impose price floors and price ceilings and shut off international trade. From this viewpoint, policies to reduce inequality may help economic output by building social support for allowing markets to operate.

The tradeoff in Figure 1(b) then flattens out in the area between points D and E, which reflects the pattern that a number of countries that provide similar levels of income to their citizens—the United States, Canada, the nations of the European Union, Japan, Australia—have different levels of inequality. The pattern suggests that countries in this range could choose a greater or a lesser degree of inequality without much impact on economic output. Only if these countries push for a much higher level of equality, like at point F, will they experience the diminished incentives that lead to lower levels of economic output. In this view, while a danger always exists that an agenda to reduce poverty or inequality can be poorly designed or pushed too far, it is also possible to discover and design policies that improve equality and do not injure incentives for economic output by very much—or even improve such incentives.

OCCUPY WALL STREET

The Occupy movement took on a life of its own over the last few months of 2011, bringing to light issues faced by many people on the lower end of the income distribution. The contents of this module indicate that there is a significant amount of income inequality in the United States. The question is: What should be done about it?

The Great Recession of 2008–2009 caused unemployment to rise and incomes to fall. Many people attribute the recession to mismanagement of the financial system by bankers and financial managers—those in the 1% of the income distribution—but those in lower quintiles bore the greater burden of the recession through unemployment. This seemed to present the picture of inequality in a different light: the group that seemed responsible for the recession was not the group that seemed to bear the burden of the decline in output. A burden shared can bring a society closer together; a burden pushed off onto others can polarize it.

On one level, the problem with trying to reduce income inequality comes down to whether you still believe in the American Dream. If you believe that one day you will have your American Dream—a large income, large house, happy family, or whatever else you would like to have in life—then you do not necessarily want to prevent anyone else from living out their dream. You certainly would not want to run the risk that someone would want to take part of your dream away from you. So there is some reluctance to engage in a redistributive policy to reduce inequality.

However, when those for whom the likelihood of living the American Dream is very small are considered, there are sound arguments in favor of trying to create greater balance. As the text indicated, a little more income equality, gained through long-term programs like increased education and job training, can increase overall economic output. Then everyone is made better off. And the 1% will not seem like such a small group any more.

Try It

Visit this page in your course online to check your understanding.

Learning Objectives

**effective income tax:** percentage of total taxes paid divided by total income

**estate tax:** a tax imposed on the value of an inheritance

**income:** a flow of money received, often measured on a monthly or an annual basis

**progressive tax system:** a tax system in which the rich pay a higher percentage of their income in taxes, rather than a higher absolute amount

**redistribution:** taking income from those with higher incomes and providing income to those with lower incomes

**wealth:** the sum of the value of all assets, including money in bank accounts, financial investments, a pension fund, and the value of a home
PUTTING IT TOGETHER: INCOME DISTRIBUTION

In this module, we learned about income distribution, poverty, and programs that make up the social safety net.

Measuring inequality involves making comparisons across the entire distribution of income, not just the poor. One way of doing this is to divide the population into groups, like quintiles, and then calculate what share of income each group receives. An alternative approach is to draw Lorenz curves, which compare the cumulative income actually received to a perfectly equal distribution of income. Income inequality in the United States increased substantially from the late 1970s and early 1980s into the 2000s. The two most common explanations that economists cite are changes in household structures that have led to more two-earner couples and single-parent families, and the effect of new information and communications technology on wages.

As the income gap has widened in recent decades, there has been increased awareness of and interest in doing more to reduce economic inequality. While most people support some sort of government assistance for those in poverty, there seems to be less consensus about limiting the incomes of those at the top of the income distribution. Pushing too aggressively for economic equality can run the risk of decreasing economic incentives. However, a moderate push for economic equality can increase economic output, both through methods like improved education and by building a base of political support for market forces.
MODULE 14: GLOBALIZATION AND TRADE

WHY IT MATTERS: GLOBALIZATION AND INTERNATIONAL TRADE

Why analyze the benefits and costs of international trade?

Over time, the world has become a smaller place. Globalization is the process by which the world, previously isolated through physical and technological distance, becomes increasingly interconnected. Globalization happens through the increase in interaction between peoples around the world that involves the sharing of ideas, cultures, goods, services and investment. The last sixty years have witnessed a huge increase in globalization, but the phenomenon has been going on for much longer. Thomas Friedman describes the current trend as the third great wave of globalization in human history.

As globalization increases over time, individuals, firms, institutions, and politicians work within and across countries to define exactly how “open” they want to be. It is natural for people to want to protect their own products through tariffs or trade restrictions, while having open access to foreign markets. But if one country protects its products, its trading partners are likely to do the same. One thing is for sure, protection and openness to international trade both have income distribution effects. The key question is who reaps the benefits and who carries the burden of the system adopted by a given country.

This module may be more important than you think. The topic is international trade and includes aspects of globalization and finance, but the theory explains every transaction we conduct. Why do people work for pay instead of growing their own food, building their own house and making their own clothes? Most people are capable of painting their own homes, yet professional painters continue to make a good living. How is international trade different from domestic trade? The answer is, “not very much, only in the details.” People buy imported goods for the same reasons they buy domestic goods. And yet we often treat foreign and domestic trade as fundamentally different. A grocery chain from a
nearby state has recently opened some stores in your neighborhood. How would you feel if the local government prohibited you from shopping at those new stores?

In this module, you will learn that just as buying from the local grocery store is better for most people than growing your own food, so international trade can add to your convenience and quality of life. And yet, most countries support some degree of protectionism, barriers to trade like tariffs, or quotas designed to “protect” domestic workers and companies.

As you proceed through this module, consider the following questions:

- What is comparative advantage?
- What are the gains from international trade?
- In what sense do barriers to trade protect American workers and companies?
- What are the costs of globalization? Are the costs worth it?

International trade and finance are often confused as being synonymous with globalization. Indeed, trade and international finance have contributed to globalization but they are not the same. Globalization is a process that widens, deepens and speeds-up interconnectedness between people, institutions, markets and nations. Trade and finance are two arteries through which the process of globalization flows. There are many ways to conceptualize globalization.

**JUST WHOSE IPHONE IS IT?**

The iPhone is a global product. Apple does not manufacture the iPhone components, nor does it assemble them. The assembly is done by Foxconn Corporation, a Taiwanese company, at its factory in Sengzhen, China. But, Samsung, the electronics firm and competitor to Apple, actually supplies many of the parts that make up an iPhone—about 26%. That means, that Samsung is both the biggest supplier and biggest competitor for Apple. Why do these two firms work together to produce the iPhone?

![Figure 1. Apple or Samsung iPhone?](https://example.com/iphone.png)

*Figure 1. Apple or Samsung iPhone? While the iPhone is readily recognized as an Apple product, 26% of the component costs in it come from components made by rival phone-maker, Samsung. In international trade, there are often “conflicts” like this as each country or company focuses on what it does best. (Credit: modification of work by Yutaka Tsutano Creative Commons)*

To understand the economic logic behind international trade, you have to accept, as these firms do, that trade is about mutually beneficial exchange. Samsung is one of the world’s largest electronics parts suppliers. Apple lets Samsung focus on making the best parts, which allows Apple to concentrate on its strength—designing elegant products that are easy to use. If each company (and by extension each country) focuses on what it does best, there will be gains for all through trade.
INTRODUCTION TO COMPARATIVE ADVANTAGE

What you'll learn to do: define and calculate comparative advantage, and understand how countries choose which goods and services to trade internationally

People trade for goods and services if they can buy them more cheaply than they could make them themselves. This is true whether you're buying produce from the grocery store or imported chocolate from another country.

We live in a global marketplace. The food on your table might include fresh fruit from Chile, cheese from France, and bottled water from Scotland. Your wireless phone might have been made in Taiwan or Korea. The clothes you wear might be designed in Italy and manufactured in China. The toys you give to a child might have come from India. The car you drive might come from Japan, Germany, or Korea. The gasoline in the tank might be refined from crude oil from Saudi Arabia, Mexico, or Nigeria. As a worker, if your job is involved with farming, machinery, airplanes, cars, scientific instruments, or many other technology-related industries, the odds are good that a hearty proportion of the sales of your employer—and hence the money that pays your salary—comes from export sales. We are all linked by international trade, and the volume of that trade has grown dramatically in the last few decades.

In this section, you will learn about the basics behind international trade, what determines the costs of imports and exports, and why it is advantageous for countries to specialize in the production of particular goods or services.
ABSOLUTE AND COMPARATIVE ADVANTAGE

Learning Objectives

- Explain absolute advantage and comparative advantage

Absolute and Comparative Advantage

The American statesman Benjamin Franklin (1706–1790) once wrote: “No nation was ever ruined by trade.” Many economists would express their attitudes toward international trade in an even more positive manner. The evidence that international trade confers overall benefits on economies is pretty strong. Trade has accompanied economic growth in the United States and around the world. Many of the national economies that have shown the most rapid growth in the last few decades—for example, Japan, South Korea, China, and India—have done so by dramatically orienting their economies toward international trade. There is no modern example of a country that has shut itself off from world trade and yet prospered. To understand the benefits of trade, or why we trade in the first place, we need to understand the concepts of comparative and absolute advantage.

In 1817, David Ricardo, a businessman, economist, and member of the British Parliament, wrote a treatise called On the Principles of Political Economy and Taxation. In this treatise, Ricardo argued that specialization and free trade benefit all trading partners, even those that may be relatively inefficient. To see what he meant, we must be able to distinguish between absolute and comparative advantage.

A country has an absolute advantage over another country if it can produce a given product using fewer resources than the other country needs to use. For example, if Canada can produce 100 pounds of beef using two ranchers, while Argentina needs three ranchers to produce 100 pounds of beef, Canada has an absolute advantage over Argentina in beef production.

Absolute advantage can be the result of a country’s natural endowment. For example, extracting oil in Saudi Arabia is pretty much just a matter of “drilling a hole.” Producing oil in other countries can require considerable exploration and costly technologies for drilling and extraction—if indeed they have any oil at all. The United States has some of the richest farmland in the world, making it easier to grow corn and wheat than in many other countries. Guatemala and Colombia have climates especially suited for growing coffee. Chile and Zambia have some of the world’s richest copper mines. As some have argued, “geography is destiny.” As a result, it should not be surprising if Chile provides copper to Guatemala, while Guatemala provides coffee to Chile. When each country has a product others need and it can be produced with fewer resources in one country over another, then it is easy to imagine all parties benefitting from trade. However, thinking about trade just in terms of geography and absolute advantage is incomplete. What happens if one country has an absolute advantage in both goods? Trade really occurs because of comparative advantage.
A country has a **comparative advantage** when it can produce a good at a lower *opportunity cost* than another country. The question each country or company should be asking when it trades is this: “What do we give up to produce this good?” For example, if Zambia produces copper, the resources it uses cannot be used to produce other goods such as corn. As a result, Zambia gives up the opportunity to produce corn. Suppose it takes 10 hours of labor to mine a ton of copper in Zambia, and 20 hours of labor to harvest a bushel of corn. This means the opportunity cost of producing a ton of copper is 2 bushels of corn. The next section develops absolute and comparative advantage in greater detail and relates them to trade.

**Try It**

Visit this page in your course online to view this presentation.

**Watch It**

Watch the following video to better understand comparative advantage.

Watch this video online: [https://youtu.be/38hvvAzgXZY](https://youtu.be/38hvvAzgXZY)

**What Happens When a Country Has an Absolute Advantage in All Goods**

What happens to the possibilities for trade if one country has an absolute advantage in everything? This is typical for high-income countries that often have well-educated workers, technologically advanced equipment, and the most up-to-date production processes. These high-income countries can produce all products with fewer resources than a low-income country. If the high-income country is more productive across the board, will there still be gains from trade? Good students of Ricardo understand that trade is about mutually beneficial exchange. Even when one country has an absolute advantage in all products, trade can still benefit both sides. This is because gains from trade come from specializing in one’s comparative advantage.

**glossary**

- **absolute advantage**: when one country can use fewer resources to produce a good compared to another country; when a country is more productive compared to another country

- **comparative advantage**: when a country can produce a good at a lower cost in terms of other goods; or, when a country has a lower opportunity cost of production

---

**Licensing & Attributions**

CC licensed content, Shared previously

- Absolute and Comparative Advantage. **Authored by**: OpenStax College. **Located at**: https://cnx.org/contents/vEmOH_p84.4A:7Nc6vlvbi/Absolute-and-Comparative-Advantages. **License**: CC BY: Attribution. **License Terms**: Download for free at http://cnx.org/content/col12498e1f-ef91-43a0-8dea-d356a0d9a82@4.4

**COMPARATIVE ADVANTAGE AND THE GAINS FROM TRADE**
Learning Objectives

- Calculate absolute and comparative advantage

Production Possibilities and Comparative Advantage

Consider the example of trade in two goods, shoes and refrigerators, between the United States and Mexico. These goods are homogeneous, meaning that consumers and producers cannot differentiate between shoes from Mexico and shoes from the U.S.; nor can they differentiate between Mexican or American refrigerators.

From Table 1, we can see that it takes four U.S. workers to produce 1,000 pairs of shoes, but it takes five Mexican workers to do so. It takes one U.S. worker to produce 1,000 refrigerators, but it takes four Mexican workers to do so. The United States has an absolute advantage in producing both shoes and refrigerators; that is, it takes fewer workers in the United States than in Mexico to produce both a given number of shoes and a given number of refrigerators.

<table>
<thead>
<tr>
<th>Country</th>
<th>Number of Workers needed to produce 1,000 units — Shoes</th>
<th>Number of Workers needed to produce 1,000 units — Refrigerators</th>
</tr>
</thead>
<tbody>
<tr>
<td>United States</td>
<td>4 workers</td>
<td>1 worker</td>
</tr>
<tr>
<td>Mexico</td>
<td>5 workers</td>
<td>4 workers</td>
</tr>
</tbody>
</table>

Absolute advantage simply compares the productivity of a worker between countries. It answers the question, “How many inputs do I need to produce shoes in Mexico?” Comparative advantage asks this same question slightly differently. Instead of comparing how many workers it takes to produce a good, it asks, “How much am I giving up to produce this good in this country?” Another way of looking at this is that comparative advantage identifies the good for which the producer’s absolute advantage is relatively larger, or where the producer's absolute productivity disadvantage is relatively smaller. The United States can produce 1,000 shoes with four-fifths as many workers as Mexico (four versus five), but it can produce 1,000 refrigerators with only one-quarter as many workers (one versus four). So, the comparative advantage of the United States, where its absolute productivity advantage is relatively greatest, lies with refrigerators, and Mexico’s comparative advantage, where its absolute productivity disadvantage is least, is in the production of shoes.

Mutually Beneficial Trade with Comparative Advantage

When nations increase production in their area of comparative advantage and trade with each other, both countries can benefit. The production possibilities frontier is a useful tool to visualize this benefit. Recall from earlier readings that the production possibilities frontier shows the maximum amount that each country can produce given its limited resources, in this case workers.

Consider a situation where the United States and Mexico each have 40 workers. For example, as Table 2 shows, if the United States divides its labor so that 40 workers are making shoes, then, since it takes four workers in the United States to make 1,000 shoes, a total of 10,000 shoes will be produced. (If four workers can make 1,000 shoes, then 40 workers will make 10,000 shoes). If the 40 workers in the United States are making refrigerators, and each worker can produce 1,000 refrigerators, then a total of 40,000 refrigerators will be produced.

<table>
<thead>
<tr>
<th>Country</th>
<th>Shoe Production — using 40 workers</th>
<th>Refrigerator Production — using 40 workers</th>
</tr>
</thead>
<tbody>
<tr>
<td>United States</td>
<td>10,000 shoes</td>
<td>or 40,000 refrigerators</td>
</tr>
</tbody>
</table>
Country | Shoe Production — using 40 workers | Refrigerator Production — using 40 workers
--- | --- | ---
Mexico | 8,000 shoes | or 10,000 refrigerators

As always, the slope of the production possibility frontier for each country is the opportunity cost of one refrigerator in terms of foregone shoe production—when labor is transferred from producing the latter to producing the former (see Figure 1).

Figure 1. Production Possibility Frontiers. (a) With 40 workers, the United States can produce either 10,000 shoes and zero refrigerators or 40,000 refrigerators and zero shoes. (b) With 40 workers, Mexico can produce a maximum of 8,000 shoes and zero refrigerators, or 10,000 refrigerators and zero shoes. All other points on the production possibility line are possible combinations of the two goods that can be produced given current resources. Point A on both graphs is where the countries start producing and consuming before trade. Point B is where they end up after trade.

Let's say that, in the situation before trade, each nation prefers to produce a combination of shoes and refrigerators that is shown at point A. Table 3 shows the output of each good for each country and the total output for the two countries.

<table>
<thead>
<tr>
<th>Country</th>
<th>Current Shoe Production</th>
<th>Current Refrigerator Production</th>
</tr>
</thead>
<tbody>
<tr>
<td>United States</td>
<td>5,000</td>
<td>20,000</td>
</tr>
<tr>
<td>Mexico</td>
<td>4,000</td>
<td>5,000</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>9,000</strong></td>
<td><strong>25,000</strong></td>
</tr>
</tbody>
</table>

Continuing with this scenario, each country transfers some amount of labor toward its area of comparative advantage. For example, the United States transfers six workers away from shoes and toward producing refrigerators. As a result, U.S. production of shoes decreases by 1,500 units (6/4 × 1,000), while its production of refrigerators increases by 6,000 (that is, 6/1 × 1,000). Mexico also moves production toward its area of comparative advantage, transferring 10 workers away from refrigerators and toward production of shoes. As a result, production of refrigerators in Mexico falls by 2,500 (10/4 × 1,000), but production of shoes increases by 2,000 pairs (10/5 × 1,000). Notice that when both countries shift production toward each of their comparative advantages (what they are relatively better at), their combined production of both goods rises, as shown in Table 4. The reduction of shoe production by 1,500 pairs in the United States is more than offset by the gain of 2,000 pairs of shoes in Mexico, while the reduction of 2,500 refrigerators in Mexico is more than offset by the additional 6,000 refrigerators produced in the United States.

<table>
<thead>
<tr>
<th>Country</th>
<th>Shoe Production</th>
<th>Refrigerator Production</th>
</tr>
</thead>
<tbody>
<tr>
<td>United States</td>
<td>3,500</td>
<td>26,000</td>
</tr>
</tbody>
</table>
This numerical example illustrates the remarkable insight of comparative advantage: even when one country has an absolute advantage in all goods and another country has an absolute disadvantage in all goods, both countries can still benefit from trade. Even though the United States has an absolute advantage in producing both refrigerators and shoes, it makes economic sense for it to specialize in the good for which it has a comparative advantage. The United States will export refrigerators and in return import shoes.

**Can a production possibility frontier be straight?**

When you first met the production possibility frontier (PPF) in an earlier module, it was drawn with an outward-bending shape. This shape illustrated that as inputs were transferred from producing one good to another—like from education to health services—there were increasing opportunity costs. In the examples in this module, the PPFs are drawn as straight lines, which means that opportunity costs are constant. When a marginal unit of labor is transferred away from growing corn and toward producing oil, the decline in the quantity of corn and the increase in the quantity of oil is always the same. In reality this is possible only if the contribution of additional workers to output did not change as the scale of production changed. The linear production possibilities frontier is a less realistic model, but a straight line simplifies calculations. It also illustrates economic themes like absolute and comparative advantage just as clearly.

**How Opportunity Cost Sets the Boundaries of Trade**

This example shows that both parties can benefit from specializing in their comparative advantages and trading. By using the opportunity costs in this example, it is possible to identify the range of possible trades that would benefit each country.

Mexico started out, before specialization and trade, producing 4,000 pairs of shoes and 5,000 refrigerators. Then, in the numerical example given, Mexico shifted production toward its comparative advantage and produced 6,000 pairs of shoes but only 2,500 refrigerators. Thus, if Mexico can export no more than 2,000 pairs of shoes (giving up 2,000 pairs of shoes) in exchange for imports of at least 2,500 refrigerators (a gain of 2,500 refrigerators), it will be able to consume more of both goods than before trade. Mexico will be unambiguously better off. Conversely, the United States started off, before specialization and trade, producing 5,000 pairs of shoes and 20,000 refrigerators. In the example, it then shifted production toward its comparative advantage, producing only 3,500 shoes but 26,000 refrigerators. If the United States can export no more than 6,000 refrigerators in exchange for imports of at least 1,500 pairs of shoes, it will be able to consume more of both goods and will be unambiguously better off.

The range of trades that can benefit both nations is shown in Table 5. For example, a trade where the U.S. exports 4,000 refrigerators to Mexico in exchange for 1,800 pairs of shoes would benefit both sides, in the sense that both countries would be able to consume more of both goods than in a world without trade.

<table>
<thead>
<tr>
<th>The U.S. economy, after specialization, will benefit if it:</th>
<th>The Mexican economy, after specialization, will benefit if it:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exports fewer than 6,000 refrigerators</td>
<td>Imports at least 2,500 refrigerators</td>
</tr>
<tr>
<td>Imports at least 1,500 pairs of shoes</td>
<td>Exports no more than 2,000 pairs of shoes</td>
</tr>
</tbody>
</table>

Trade allows each country to take advantage of lower opportunity costs in the other country. If Mexico wants to produce more refrigerators without trade, it must face its domestic opportunity costs and reduce shoe production. If Mexico, instead, produces more shoes and then trades for refrigerators made in the United States, where the opportunity cost of producing refrigerators is lower, Mexico can in effect take advantage of the lower opportunity cost of refrigerators in the United States. Conversely, when the United States specializes in its comparative advantage of refrigerator production.
and trades for shoes produced in Mexico, international trade allows the United States to take advantage of the lower opportunity cost of shoe production in Mexico.

The theory of comparative advantage explains why countries trade: they have different comparative advantages. It shows that the gains from international trade result from pursuing comparative advantage and producing at a lower opportunity cost. The following feature shows how to calculate absolute and comparative advantage and the way to apply them to a country's production.

### Calculating Absolute and Comparative Advantage

In Canada a worker can produce 20 barrels of oil or 40 tons of lumber. In Venezuela, a worker can produce 60 barrels of oil or 30 tons of lumber.

<table>
<thead>
<tr>
<th>Country</th>
<th>Oil (barrels)</th>
<th>Lumber (tons)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Canada</td>
<td>20</td>
<td>or 40</td>
</tr>
<tr>
<td>Venezuela</td>
<td>60</td>
<td>or 30</td>
</tr>
</tbody>
</table>

1. Who has the absolute advantage in the production of oil or lumber? How can you tell?
2. Which country has a comparative advantage in the production of oil?
3. Which country has a comparative advantage in producing lumber?
4. In this example, is absolute advantage the same as comparative advantage, or not?
5. In what product should Canada specialize? In what product should Venezuela specialize?

**Step 1.** Make a table like Table 6.

**Step 2.** To calculate absolute advantage, look at the larger of the numbers for each product. One worker in Canada can produce more lumber (40 tons versus 30 tons), so Canada has the absolute advantage in lumber. One worker in Venezuela can produce 60 barrels of oil compared to a worker in Canada who can produce only 20.

**Step 3.** To calculate comparative advantage, find the opportunity cost of producing one barrel of oil in both countries. The country with the lowest opportunity cost has the comparative advantage. With the same labor time, Canada can produce either 20 barrels of oil or 40 tons of lumber. So in effect, 20 barrels of oil is equivalent to 40 tons of lumber: 20 oil = 40 lumber. Divide both sides of the equation by 20 to calculate the opportunity cost of one barrel of oil in Canada. 20/20 oil = 40/20 lumber. 1 oil = 2 lumber. To produce one additional barrel of oil in Canada has an opportunity cost of 2 lumber. Calculate the same way for Venezuela: 60 oil = 30 lumber. Divide both sides of the equation by 60. One oil in Venezuela has an opportunity cost of 1/3 lumber. Because 1/3 lumber < 2 lumber, Venezuela has the comparative advantage in producing oil.

**Step 4.** Calculate the opportunity cost of one lumber by reversing the numbers, with lumber on the left side of the equation. In Canada, 40 lumber is equivalent in labor time to 20 barrels of oil: 40 lumber = 20 oil. Divide each side of the equation by 40. The opportunity cost of one lumber is 1/2 oil. In Venezuela, the equivalent labor time will produce 30 lumber or 60 oil: 30 lumber = 60 oil. Divide each side by 30. One lumber has an opportunity cost of two oil. Canada has the lower opportunity cost in producing lumber.

**Step 5.** In this example, absolute advantage is the same as comparative advantage. Canada has the absolute and comparative advantage in lumber; Venezuela has the absolute and comparative advantage in oil.

**Step 6.** Canada should specialize in what it has a relative lower opportunity cost, which is lumber, and Venezuela should specialize in oil. Canada will be exporting lumber and importing oil, and Venezuela will be exporting oil and importing lumber.

### Try It

Visit this page in your course online to view this presentation.
Trade and Incomes

Incomes depend on labor productivity. A country with an absolute advantage in some product has higher labor productivity than another country does in the production of that product. If a country has an absolute advantage in producing both goods, it has higher labor productivity in both and its workers will earn higher incomes than those in the other country. Thus, the average income in a country depends on its average labor productivity. Now consider comparative advantage. If a country specializes production in the product in which it has a comparative advantage, it raises its average labor productivity and raises its average income. Thus, comparative advantage is more important than absolute advantage in understanding which country should trade which product in order to maximize the standard of living in both countries.

Watch It

Watch this video to review the ways that comparative advantage benefits all the parties involved.
Watch this video online: https://youtu.be/4rUfoU04QJM
For additional practice and review using numbers, watch this video from ACDC economics.

LEARN BY DOING: COMPARATIVE ADVANTAGE AND THE GAINS FROM TRADE

Try It

These questions allow you to get as much practice as you need, as you can click the link at the top of the first question (“Try another version of these questions”) to get a new set of questions. Practice until you feel comfortable doing the questions.
Visit this page in your course online to practice before taking the quiz.
INTRA-INDUSTRY TRADE

Learning Objectives

- Identify at least two advantages of intra-industry trading
- Explain the relationship between economies of scale and intra-industry trade

Absolute and comparative advantages explain a great deal about global trading patterns. For example, they help to explain the patterns that we noted at the start of this module, like why you may be eating fresh fruit from Chile or Mexico, or why lower productivity regions like Africa and Latin America are able to sell a substantial proportion of their exports to higher productivity regions like the European Union and North America. Comparative advantage, however, at least at first glance, does not seem especially well-suited to explain other common patterns of international trade.

The Prevalence of Intra-industry Trade between Similar Economies

The theory of comparative advantage suggests that trade should happen between economies with large differences in opportunity costs of production. Roughly half of all world trade involves shipping goods between the fairly similar high-income economies of the United States, Canada, the European Union, Japan, Mexico, and China (see Table 1).


<table>
<thead>
<tr>
<th>Country</th>
<th>U.S. Exports Go to ...</th>
<th>U.S. Imports Come from ...</th>
</tr>
</thead>
<tbody>
<tr>
<td>European Union</td>
<td>19.0%</td>
<td>21.0%</td>
</tr>
</tbody>
</table>
Moreover, the theory of comparative advantage suggests that each economy should specialize to a degree in certain products, and then exchange those products. A high proportion of trade, however, is **intra-industry trade**—that is, trade of goods within the same industry from one country to another. For example, the United States produces and exports autos and imports autos. Table 2 shows some of the largest categories of U.S. exports and imports. In all of these categories, the United States is both a substantial exporter and a substantial importer of goods from the same industry. In 2014, according to the Bureau of Economic Analysis, the United States exported $146 billion worth of autos, and imported $327 billion worth of autos. About 60% of U.S. trade and 60% of European trade is intra-industry trade.

<table>
<thead>
<tr>
<th>Country</th>
<th>U.S. Exports Go to ...</th>
<th>U.S. Imports Come from ...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Canada</td>
<td>22.0%</td>
<td>14.0%</td>
</tr>
<tr>
<td>Japan</td>
<td>4.0%</td>
<td>6.0%</td>
</tr>
<tr>
<td>Mexico</td>
<td>15.0%</td>
<td>13.0%</td>
</tr>
<tr>
<td>China</td>
<td>8.0%</td>
<td>20.0%</td>
</tr>
</tbody>
</table>

Table 2. Some Intra-Industry U.S. Exports and Imports in 2014 (Source: [http://www.bea.gov/newsreleases/international/trade/tradnewsrelease.htm](http://www.bea.gov/newsreleases/international/trade/tradnewsrelease.htm))

Why do similar high-income economies engage in intra-industry trade? What can be the economic benefit of having workers of fairly similar skills making cars, computers, machinery and other products which are then shipped across the oceans to and from the United States, the European Union, and Japan? There are two reasons: (1) The division of labor leads to learning, innovation, and unique skills; and (2) economies of scale.

**Try It**

Visit this page in your course online to view this presentation.

**Watch It**

Watch the selected portion of this video to see examples of intra-industry trade. The video introduces the Grubel-Lloyd Index, which assigns a numerical value for measuring and comparing intra-industry trade.

Watch this video online: [https://youtu.be/DUmgU_F3lmk](https://youtu.be/DUmgU_F3lmk)

**Gains from Specialization and Learning**

Consider the category of machinery, where the U.S. economy has considerable intra-industry trade. Machinery comes in many varieties, so the United States may be exporting machinery for manufacturing with wood, but importing machinery for photographic processing. The underlying reason why a country like the United States, Japan, or Germany produces one kind of machinery rather than another is usually not related to U.S., German, or Japanese firms and workers having
generally higher or lower skills. It is just that, in working on very specific and particular products, firms in certain countries develop unique and different skills.

Specialization in the world economy can be very finely split. In fact, recent years have seen a trend in international trade called splitting up the value chain. The value chain describes how a good is produced in stages. As indicated in the beginning of the module, the production of the iPhone involves the design and engineering of the phone in the United States, parts supplied from Korea, the assembly of the parts in China, and the advertising and marketing done in the United States. Thanks in large part to improvements in communication technology, sharing information, and transportation, it has become easier to split up the value chain.

Instead of production in a single large factory, all of these steps can be split up among different firms operating in different places and even different countries. Because firms split up the value chain, international trade often does not involve whole finished products like automobiles or refrigerators being traded between nations. Instead, it involves shipping more specialized goods like, say, automobile dashboards or the shelving that fits inside refrigerators. Intra-industry trade between similar countries produces economic gains because it allows workers and firms to learn and innovate on particular products—and often to focus on very particular parts of the value chain.

Economies of Scale, Competition, Variety

A second broad reason that intra-industry trade between similar nations produces economic gains involves economies of scale. The concept of economies of scale means that as the scale of output goes up, average costs of production decline—at least up to a point. Figure 2 illustrates economies of scale for a plant producing toaster ovens. The horizontal axis of the figure shows the quantity of production by a certain firm or at a certain manufacturing plant. The vertical axis measures the average cost of production. Production plant S produces a small level of output at 30 units and has an average cost of production of $30 per toaster oven. Plant M produces at a medium level of output at 50 units, and has an average cost of production of $20 per toaster oven. Plant L produces 150 units of output with an average cost of production of only $10 per toaster oven. Although plant V can produce 200 units of output, it still has the same unit cost as Plant L.

In this example, a small or medium plant, like S or M, will not be able to compete in the market with a large or a very large plant like L or V, because the firm that operates L or V will be able to produce and sell their output at a lower price. In this example, economies of scale operate up to point L, but beyond point L to V, the additional scale of production does not continue to reduce average costs of production.

Figure 1. Economies of Scale Production. Plant S, has an average cost of production of $30 per toaster oven. Production plant M has an average cost of production of $20 per toaster oven. Production plant L has an average cost of production of only $10 per toaster oven. Production plant V would still have an average cost of production of $10 per toaster oven. Thus, production plant M can produce toaster ovens more cheaply than plant S because of economies of scale, and plants L or V can produce more cheaply than S or M because of economies of scale. However, the economies of scale end at an output level of 150. Plant V, despite being larger, cannot produce more cheaply on average than plant L.
The concept of economies of scale becomes especially relevant to international trade when it enables one or two large producers to supply the entire country. For example, a single large automobile factory could probably supply all the cars purchased in a smaller economy like the United Kingdom or Belgium in a given year. However, if a country has only one or two large factories producing cars, and no international trade, then consumers in that country would have relatively little choice between kinds of cars (other than the color of the paint and other nonessential options). Little or no competition will exist between different car manufacturers.

International trade provides a way to combine the lower average production costs that come from economies of scale and still have competition and variety for consumers. Large automobile factories in different countries can make and sell their products around the world. If the U.S. automobile market was made up of only General Motors, Ford, and Chrysler, the level of competition and consumer choice would be quite a lot lower than when U.S. carmakers must face competition from Toyota, Honda, Suzuki, Fiat, Mitsubishi, Nissan, Volkswagen, Kia, Hyundai, BMW, Subaru, and others. Greater competition brings with it innovation and responsiveness to what consumers want. America’s car producers make far better cars now than they did several decades ago, and much of the reason is competitive pressure, especially from East Asian and European carmakers.

Try It
Visit this page in your course online to view this presentation.

It is Apple’s (Global) iPhone

Apple Corporation uses a global platform to produce the iPhone. Now that you understand the concept of comparative advantage, you can see why the engineering and design of the iPhone is done in the United States. The United States has built up a comparative advantage over the years in designing and marketing products, and sacrifices fewer resources to design high-tech devices relative to other countries. China has a comparative advantage in assembling the phone due to its large skilled labor force. Korea has a comparative advantage in producing components. Korea focuses its production by increasing its scale, learning better ways to produce screens and computer chips, and uses innovation to lower average costs of production. Apple, in turn, benefits because it can purchase these quality products at lower prices. Put the global assembly line together and you have the device with which we are all so familiar.

Dynamic Comparative Advantage

The sources of gains from intra-industry trade between similar economies—namely, the learning that comes from a high degree of specialization and splitting up the value chain and from economies of scale—do not contradict the earlier theory of comparative advantage. Instead, they help to broaden the concept.

In intra-industry trade, the level of worker productivity is not determined by climate or geography. It is not even determined by the general level of education or skill. Instead, the level of worker productivity is determined by how firms engage in specific learning about specialized products, including taking advantage of economies of scale. In this vision, comparative advantage can be dynamic—that is, it can evolve and change over time as new skills are developed and as the value chain is split up in new ways. This line of thinking also suggests that countries are not destined to have the same comparative advantage forever, but must instead be flexible in response to ongoing changes in comparative advantage.

Glossary

**Grubel-Lloyd Index**: measure of interindustry trade in some industry; values closer to zero indicate low interindustry trade; values closer to one indicate high interindustry trade

**intra-industry trade**: international trade of goods within the same industry

**splitting up the value chain**: conducting different stages of production of a good in different geographic locations
DEMAND AND SUPPLY ANALYSIS OF INTERNATIONAL TRADE

Learning Objectives

- Use supply and demand to explain the gains from trade

Demand and Supply Analysis of International Trade

We can use the theory of supply and demand to further understand the benefits of international trade. Consider two countries, Brazil and the United States, who produce sugar. Each country has a domestic supply and demand for sugar, as detailed in Table 1 and illustrated in Figure 2. In Brazil, without trade, the equilibrium price of sugar is 12 cents per pound and the equilibrium output is 30 tons. When there is no trade in the United States, the equilibrium price of sugar is 24 cents per pound and the equilibrium quantity is 80 tons. These equilibrium points are labeled with the point E. Notice that in this set-up, Brazil is the low-cost provider of sugar and has the cost-advantage.

<table>
<thead>
<tr>
<th>Price</th>
<th>Brazil: Quantity Supplied (tons)</th>
<th>Brazil: Quantity Demanded (tons)</th>
<th>U.S.: Quantity Supplied (tons)</th>
<th>U.S.: Quantity Demanded (tons)</th>
</tr>
</thead>
<tbody>
<tr>
<td>8 cents</td>
<td>20</td>
<td>35</td>
<td>60</td>
<td>100</td>
</tr>
<tr>
<td>12 cents</td>
<td>30</td>
<td>30</td>
<td>66</td>
<td>93</td>
</tr>
<tr>
<td>14 cents</td>
<td>35</td>
<td>28</td>
<td>69</td>
<td>90</td>
</tr>
<tr>
<td>16 cents</td>
<td>40</td>
<td>25</td>
<td>72</td>
<td>87</td>
</tr>
<tr>
<td>20 cents</td>
<td>45</td>
<td>21</td>
<td>76</td>
<td>83</td>
</tr>
<tr>
<td>24 cents</td>
<td>50</td>
<td>18</td>
<td>80</td>
<td>80</td>
</tr>
</tbody>
</table>
### Table 1. The Sugar Trade between Brazil and the United States

<table>
<thead>
<tr>
<th>Price</th>
<th>Brazil: Quantity Supplied (tons)</th>
<th>Brazil: Quantity Demanded (tons)</th>
<th>U.S.: Quantity Supplied (tons)</th>
<th>U.S.: Quantity Demanded (tons)</th>
</tr>
</thead>
<tbody>
<tr>
<td>28 cents</td>
<td>55</td>
<td>15</td>
<td>82</td>
<td>78</td>
</tr>
</tbody>
</table>

#### Figure 1. Free trade results in gains from trade. Total surplus increases in both countries. However, there are clear income distribution effects.

If international trade between Brazil and the United States now becomes possible, profit-seeking firms will spot an opportunity: buy sugar cheaply in Brazil, and sell it at a higher price in the United States. As sugar is shipped from Brazil to the United States, the quantity of sugar produced in Brazil will be greater than Brazilian consumption (with the extra production being exported), and the amount produced in the United States will be less than the amount of U.S. consumption (with the extra consumption being imported). Exports to the United States will reduce the supply of sugar in Brazil, raising its price. Imports into the United States will increase the supply of sugar, lowering its price. When the price of sugar is the same in both countries, there is no incentive to trade further. As Figure 1 shows, the equilibrium with trade occurs at a price of 16 cents per pound. At that price, the sugar farmers of Brazil supply a quantity of 40 tons, while the consumers of Brazil buy only 25 tons.

The extra 15 tons of sugar production, shown by the horizontal gap between the demand curve and the supply curve in Brazil, is exported to the United States. In the United States, at a price of 16 cents, the farmers produce a quantity of 72 tons and consumers demand a quantity of 87 tons. The excess demand of 15 tons by American consumers, shown by the horizontal gap between demand and domestic supply at the price of 16 cents, is supplied by imported sugar.

Free trade typically results in income distribution effects, but the key is to recognize the overall gains from trade, as shown in Figure 1. Building on the concepts you have already learned about supply and demand and consumer and producer surplus, Figure 1(a) shows that producers in Brazil gain by selling more sugar at a higher price, while Figure 1(b) shows consumers in the United States benefit from the lower price and greater availability of sugar. Consumers in Brazil are worse off (compare their no-trade consumer surplus with the free-trade consumer surplus) and U.S. producers of sugar are worse off. There are gains from trade—an increase in social surplus in each country. That is, both the United States and Brazil are better off than they would be without trade.
Figure 2. When there is free trade, the equilibrium is at point A. When there is no trade, the equilibrium is at point E.

Try It

Visit this page in your course online to view this presentation.

LEARN BY DOING: DEMAND AND SUPPLY ANALYSIS OF INTERNATIONAL TRADE

Try It

These questions allow you to get as much practice as you need, as you can click the link at the top of the first question ("Try another version of these questions") to get a new set of questions. Practice until you feel comfortable doing the questions.

Visit this page in your course online to practice before taking the quiz.

Try It

These questions allow you to get as much practice as you need, as you can click the link at the top of the first question ("Try another version of these questions") to get a new set of questions. Practice until you feel comfortable doing the questions.
INTRODUCTION TO THE TRADE BARRIERS AND PROTECTIONISM

What you'll learn to do: explain how barriers to trade (like tariffs, quotas and non-tariff barriers) affect businesses, consumers and workers in the economy

We have seen that international trades raises the standard of living for participating countries. Indeed, free trade maximizes the gains from international trade. While each country is better off through international trade (or more precisely, the average resident is better off), that doesn't mean that all individuals are better off. Free trade is a policy and like every policy, there are winners and losers. The winners are consumers and workers, managers and owners of firms that produce goods whose demand increases through international trade. The losers are workers, managers and owners of firms whose demand decreases as a result of international trade; that is, firms who produce substitutes for imports.

Protectionism is an attempt to mitigate the harm done by international trade. The most prominent argument for tariffs, quotas and other barriers to trade is to protect jobs and incomes that otherwise would be at risk from foreign imports. This section will explore these issues.
PROTECTIONISM

Learning Objectives

- Describe why governments may justify protectionist policies
- Explain and give examples of trade barriers, including quotas, tariffs, and nontariff barriers

Protectionism

Globalization has brought fear of loss of jobs and loss of income, which are often described as the “race to the bottom,” as industrialized countries are thought to have to reduce wages to be competitive with those in the developing world. Globalization has also spawned fears about loss of culture. Many countries worry about their cultures being overwhelmed by that of the United States. France is a good example. Others fear replacement of their cultures by that of Western nations (e.g., some Islamic states). Countries also fear the loss of national sovereignty as they become part of supranational entities, like the European Union or the International Monetary Fund. And yet, history shows that globalization has corresponded to higher national incomes and increased opportunities. How can these conflicting views be reconciled?

When a government legislates policies to reduce or block international trade it is engaging in protectionism. Protectionist policies often seek to shield domestic producers and domestic workers from foreign competition. The Trump Administration’s tariffs on steel and aluminum in 2018 are a recent example.

Watch It

A government may justify protectionist policies for one of the following reasons, which are outlined in this video:

1. Protect domestic jobs.
2. Level the playing field.
3. Raise additional revenue for the domestic government.
5. Infant industries—protect new industries until they are more mature
6. Promote exports.

Watch this video online: https://youtu.be/Y2X3KPilAt0

HOW DOES THE UNITED STATES REALLY FEEL ABOUT EXPANDING TRADE?

How do people around the world feel about expanding trade between nations? In summer 2007, the Pew Foundation surveyed 45,000 people in 47 countries. One of the questions asked about opinions on growing trade ties between
countries. Table 20.3 shows the percentages who answered either “very good” or “somewhat good” for some of the countries surveyed.

For those who think of the United States as the world’s leading supporter of expanding trade, the survey results may be perplexing. When adding up the shares of those who say that growing trade ties between countries is “very good” or “somewhat good,” Americans had the least favorable attitude toward increasing globalization, while the Chinese and South Africans ranked highest. In fact, among the 47 countries surveyed, the United States ranked by far the lowest on this measure, followed by Egypt, Italy, and Argentina.

<table>
<thead>
<tr>
<th>Country</th>
<th>Very Good</th>
<th>Somewhat Good</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>China</td>
<td>38%</td>
<td>53%</td>
<td>91%</td>
</tr>
<tr>
<td>South Africa</td>
<td>42%</td>
<td>43%</td>
<td>87%</td>
</tr>
<tr>
<td>South Korea</td>
<td>24%</td>
<td>62%</td>
<td>86%</td>
</tr>
<tr>
<td>Germany</td>
<td>30%</td>
<td>55%</td>
<td>85%</td>
</tr>
<tr>
<td>Canada</td>
<td>29%</td>
<td>53%</td>
<td>82%</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>28%</td>
<td>50%</td>
<td>78%</td>
</tr>
<tr>
<td>Mexico</td>
<td>22%</td>
<td>55%</td>
<td>77%</td>
</tr>
<tr>
<td>Brazil</td>
<td>13%</td>
<td>59%</td>
<td>72%</td>
</tr>
<tr>
<td>Japan</td>
<td>17%</td>
<td>55%</td>
<td>72%</td>
</tr>
<tr>
<td>United States</td>
<td>14%</td>
<td>45%</td>
<td>59%</td>
</tr>
</tbody>
</table>

Table 1. The Status of Growing Trade Ties between Countries. (Source: http://www.pewglobal.org/files/pdf/258.pdf)

Try It

Visit this page in your course online to view this presentation.

Barriers to Trade

Protectionism takes three main forms: tariffs, import quotas, and nontariff barriers. Tariffs are taxes that a government imposes on imported goods and services. This makes imports more expensive for consumers, discouraging purchases of imports in favor of domestic substitutes. For example, in recent years large, flat-screen televisions imported to the U.S. from China have faced a 5% tariff rate.

Another way to control trade is through import quotas, which are numerical limitations on the quantity of products that a country can import. For instance, during the early 1980s, the Reagan Administration imposed a quota on the import of Japanese automobiles. In the 1970s, many developed countries, including the United States, found themselves with declining textile industries. Textile production does not require highly skilled workers, so producers were able to set up lower-cost factories in developing countries. In order to “manage” this loss of jobs and income, the developed countries established an international Multifiber Agreement that essentially divided the market for textile exports between importers and the remaining domestic producers. The agreement, which ran from 1974 to 2004, specified the exact quota of textile imports that each developed country would accept from each low-income country. A similar story exists for sugar imports into the United States, which are still governed by quotas.
**Nontariff barriers** are all the other ways that a nation can draw up rules, regulations, inspections, and paperwork to make it more costly or difficult to import products. A rule requiring certain safety standards can limit imports just as effectively as high tariffs or low import quotas, for instance. There are also nontariff barriers in the form of "rules-of-origin" regulations; these rules describe the “Made in Country X” label as the one in which the last substantial change in the product took place. A manufacturer wishing to evade import restrictions may try to change the production process so that the last big change in the product happens in his or her own country. For example, certain textiles are made in the United States, shipped to other countries, combined with textiles made in those other countries to make apparel—and then re-exported back to the United States for a final assembly, to escape paying tariffs or to obtain a “Made in the USA” label.

Despite import quotas, tariffs, and nontariff barriers, the share of apparel sold in the United States that is imported rose from about half in 1999 to about three-quarters today. The U.S. Bureau of Labor Statistics (BLS), estimated the number of U.S. jobs in textiles and apparel fell from 666,360 in 2007 to 385,240 in 2012, a 42% decline. Even more U.S. textile industry jobs would have been lost without tariffs. However, domestic jobs that are saved by import quotas come at a cost. Because textile and apparel protectionism adds to the costs of imports, consumers end up paying billions of dollars more for clothing each year. Some of those “consumers” are domestic producers of other goods, like motor vehicles, for example. Higher prices for steel and aluminum increase the cost of producing motor vehicles, making them harder to sell domestically and internationally. Thus, it’s not clear that protectionism saves domestic jobs or incomes.

When the United States eliminates trade barriers in one area, consumers spend the money they save on that product elsewhere in the economy. Thus, while eliminating trade barriers in one sector of the economy will likely result in some job loss in that sector, consumers will spend the resulting savings in other sectors of the economy and hence increase the number of jobs in those other sectors. Of course, workers in some of the poorest countries of the world who would otherwise have jobs producing textiles, would gain considerably if the United States reduced its barriers to trade in textiles. That said, there are good reasons to be wary about reducing barriers to trade. The 2012 and 2013 Bangladeshi fires in textile factories, which resulted in a horrific loss of life, present complications that our simplified analysis in the chapter will not capture.

**Realizing the compromises between nations that come about due to trade policy, many countries came together in 1947 to form the General Agreement on Tariffs and Trade (GATT). (We’ll cover the GATT in more detail later in the module.) This agreement has since been superseded by the World Trade Organization (WTO), whose membership includes about 150 nations and most of the world’s economies. It is the primary international mechanism through which nations negotiate their trade rules—including rules about tariffs, quotas, and nontariff barriers. The next section examines the results of such protectionism.**

**Glossary**

- **globalization**: the increase in interaction between peoples around the world that involves the sharing of ideas, cultures, goods, services and investment
- **import quotas**: numerical limits on the quantity of products that a country can import
- **nontariff barriers**: ways a nation can draw up rules, regulations, inspections, and paperwork to make it more costly or difficult to import products
**THE TRADEOFFS OF INTERNATIONAL TRADE**

**Learning Objectives**

- Analyze the tradeoffs of trade policy, identifying at least two benefits from reducing trade barriers

The Tradeoffs of International Trade

Economists readily acknowledge that international trade is not all sunshine, roses, and happy endings. Over time, the average person gains from international trade, both as a worker who has greater productivity and higher wages because of the benefits of specialization and comparative advantage, and as a consumer who can benefit from shopping all over the world for a greater variety of quality products at attractive prices. The “average person,” however, is hypothetical, not real—representing a mix of those who have done very well, those who have done all right, and those who have done poorly. It is a legitimate concern of public policy to focus not just on the average or on the success stories, but also on those who have not been so fortunate. Workers in other countries, the environment, and prospects for new industries and materials that might be of key importance to the national economy are also all legitimate issues.

The common belief among economists is that it is better to embrace the gains from trade, and then deal with the costs and tradeoffs with other policy tools, for example, retraining workers who lose their jobs to imports, than it is to cut off trade to avoid the costs and tradeoffs.

To gain a better intuitive understanding for this argument, consider a hypothetical American company called Technotron. Technotron invents a new scientific technology that allows the firm to increase the output and quality of its goods with a smaller number of workers at a lower cost. As a result of this technology, other U.S. firms in this industry will lose money and will also have to lay off workers—and some of the competing firms will even go bankrupt. Should the United States government protect the existing firms and their employees by making it illegal for Technotron to use its new technology?

Most people who live in market-oriented economies would oppose trying to block better products that lower the cost of services. Certainly, there is a case for society providing temporary support and assistance for those who find
themselves without work. Many would argue for government support of programs that encourage retraining and acquiring additional skills. Government might also support research and development efforts, so that other firms may find ways of outdoing Technotron. Blocking the new technology altogether, however, seems like a mistake. After all, few people would advocate giving up electricity because it caused so much disruption to the kerosene and candle business. Few would suggest holding back on improvements in medical technology because they might cause companies selling leeches and snake oil to lose money. In short, most people view disruptions due to technological change as a necessary cost that is worth bearing.

Now, imagine that Technotron’s new “technology” is as simple as this: the company imports what it sells from another country. In other words, think of foreign trade as a type of innovative technology. The objective situation is now exactly the same as before. Because of Technotron’s new technology—which in this case is importing goods from another county—other firms in this industry will lose money and lay off workers. Just as it would have been inappropriate and ultimately foolish to respond to the disruptions of new scientific technology by trying to shut it down, it would be inappropriate and ultimately foolish to respond to the disruptions of international trade by trying to restrict trade.

Some workers and firms will suffer because of international trade. In a living, breathing market-oriented economy, some workers and firms will always be experiencing disruptions, for a wide variety of reasons. Corporate management can be better or worse. Workers for a certain firm can be more productive or less. Tough domestic competitors can create just as much disruption as tough foreign competitors. Sometimes a new product is a hit with consumers; sometimes it is a flop. Sometimes a company is blessed by a run of good luck or stricken with a run of bad luck. For some firms, international trade will offer great opportunities for expanding productivity and jobs; for other firms, trade will impose stress and pain. The disruption caused by international trade is not fundamentally different from all the other disruptions caused by the other workings of a market economy.

In other words, the economic analysis of free trade does not rely on a belief that foreign trade is not disruptive or does not pose tradeoffs; indeed, the story of Technotron begins with a particular disruptive market change—a new technology—that causes real tradeoffs. In thinking about the disruptions of foreign trade, or any of the other possible costs and tradeoffs of foreign trade discussed in this module, the best public policy solutions typically do not involve protectionism, but instead involve finding ways for public policy to address the particular issues, while still allowing the benefits of international trade to occur.

Low-income countries benefit more from trade than high-income countries do. In some ways, the giant U.S. economy has less need for international trade, because it can already take advantage of internal trade within its economy. However, many smaller national economies around the world, in regions like Latin America, Africa, the Middle East, and Asia, have much more limited possibilities for trade inside their countries or their immediate regions. Without international trade, they may have little ability to benefit from comparative advantage, slicing up the value chain, or economies of scale. Moreover, smaller economies often have fewer competitive firms making goods within their economy, and thus firms have less pressure from other firms to provide the goods and prices that consumers want.

The economic gains from expanding international trade are measured in hundreds of billions of dollars, and the gains from international trade as a whole probably reach well into the trillions of dollars. The potential for gains from trade may be especially high among the smaller and lower-income countries of the world.

Try It
Visit this page in your course online to view this presentation.

WHAT’S THE DOWNSIDE OF PROTECTION?

Flat-panel displays, the displays for laptop computers, tablets, and flat screen televisions, are an example of such an enduring principle. In the early 1990s, the vast majority of flat-panel displays used in U.S.-manufactured laptops were imported, primarily from Japan. The small but politically powerful U.S. flat-panel-display industry filed a dumping complaint with the Commerce Department. They argued that Japanese firms were selling displays at “less than fair value,” which made it difficult for U.S. firms to compete. After a preliminary determination by the Commerce Department that the Japanese firms were dumping, the U.S. International Trade Commission imposed a 63% dumping margin (or tax) on the import of flat-panel displays. Was this a successful exercise of U.S. trade policy? Flat-panel displays make up a significant portion of the cost of producing laptop computers—as much as 50%. Therefore, the antidumping tax would substantially increase the cost, and thus the price, of U.S.-manufactured
laptops. As a result of the ITC’s decision, Apple moved its domestic manufacturing plant for Macintosh computers to Ireland (where it had an existing plant). Toshiba shut down its U.S. manufacturing plant for laptops. And IBM cancelled plans to open a laptop manufacturing plant in North Carolina, instead deciding to expand production at its plant in Japan. In this case, rather than having the desired effect of protecting U.S. interests and giving domestic manufacturing an advantage over items manufactured elsewhere, it had the unintended effect of driving the manufacturing completely out of the country. Many people lost their jobs and most flat-panel display production now occurs in countries other than the United States.

From Interpersonal to International Trade

Most people find it easy to believe that they, personally, would not be better off if they tried to grow and process all of their own food, to make all of their own clothes, to build their own cars and houses from scratch, and so on. Instead, we all benefit from living in economies where people and firms can specialize and trade with each other.

The benefits of trade do not stop at national boundaries, either. The division of labor could increase output for three reasons: (1) workers with different characteristics can specialize in the types of production where they have a comparative advantage; (2) firms and workers who specialize in a certain product become more productive with learning and practice; and (3) economies of scale. These three reasons apply from the individual and community level right up to the international level. If it makes sense to you that interpersonal, intercommunity, and interstate trade offer economic gains, it should make sense that international trade offers gains, too.

International trade currently involves about $20 trillion worth of goods and services moving around the globe. Any economic force of that size, even if it confers overall benefits, is certain to cause disruption and controversy.

Glossary

disruptive market change: innovative new product or production technology which disrupts the status quo in a market, leading the innovators to earn more income and profits and the other firms to lose income and profits, unless they can come up with their own innovations
What you'll learn to do: differentiate between alternative international trade regimes and how they impact global trade

Every country must adopt a policy on international trade. Free trade is one option, as is autarky (no international trade), and everything in between. Political pressures to protect special interests collide with those promoting the benefits of free trade. The result is a country’s trade policy, which is often a compromise between competing politics.

THE ROLE OF THE GATT IN REDUCING BARRIERS TO TRADE

Learning Objectives

- Explain the origin and role of the General Agreement on Tariffs and Trade (GATT)

Reducing Barriers to Trade

As you know, tariffs are taxes that governments place on imported goods for a variety of reasons. Some of these reasons include protecting sensitive industries, for humanitarian reasons, and protecting against dumping. Traditionally, tariffs were often used as a political tool to protect certain vested economic, social, and cultural interests.

At the beginning of the Great Depression in 1930, the U.S. Congress passed the Smoot-Hawley Tariff Act to protect American jobs and industries from foreign competition. This act raised U.S. tariffs on dutiable imports to nearly 60%. U.S. trading partners retaliated by raising their own tariffs on U.S. exports, with the result that international trade between the warring nations declined by half. The consensus among economists is that the Smoot-Hawley tariffs contributed significantly to the depth and length of the Great Depression.
At the end of World War II, there was a consensus that tariffs were too high worldwide, and that tariff reductions could stimulate international trade and return the world to a thriving, peacetime economy. In the years after the Great Depression and World War II, there was a worldwide push to build institutions that would tie the nations of the world together. The United Nations officially came into existence in 1945. The World Bank, which assists the poorest people in the world, and the International Monetary Fund, which addresses issues raised by international financial transactions, were both created in 1946. The third planned organization was to be an International Trade Organization, which would manage international trade. The United Nations was unable to agree to this. Instead, the General Agreement on Tariffs and Trade (GATT), was established in 1947 to provide a forum in which nations could come together to negotiate reductions in tariffs and other barriers to trade. In 1995, the GATT was transformed into the World Trade Organization (WTO).

The GATT process was to negotiate an agreement to reduce barriers to trade, sign that agreement, pause for a while, and then start negotiating the next agreement. The rounds of talks in the GATT, and now the WTO, are shown in Table 1. Notice that the early rounds of GATT talks took a relatively short time, included a small number of countries, and focused almost entirely on reducing tariffs. Since the 1970s, however, rounds of trade talks have taken years, included a large number of countries, and an ever-broadening range of issues.

<table>
<thead>
<tr>
<th>Year</th>
<th>Place or Name of Round</th>
<th>Main Subjects</th>
<th>Number of Countries Involved</th>
</tr>
</thead>
<tbody>
<tr>
<td>1947</td>
<td>Geneva</td>
<td>Tariff reduction</td>
<td>23</td>
</tr>
<tr>
<td>1949</td>
<td>Annecy</td>
<td>Tariff reduction</td>
<td>13</td>
</tr>
<tr>
<td>1951</td>
<td>Torquay</td>
<td>Tariff reduction</td>
<td>38</td>
</tr>
<tr>
<td>1956</td>
<td>Geneva</td>
<td>Tariff reduction</td>
<td>26</td>
</tr>
<tr>
<td>1960–61</td>
<td>Dillon round</td>
<td>Tariff reduction</td>
<td>26</td>
</tr>
<tr>
<td>1964–67</td>
<td>Kennedy round</td>
<td>Tariffs, anti-dumping measures</td>
<td>62</td>
</tr>
<tr>
<td>1973–79</td>
<td>Tokyo round</td>
<td>Tariffs, nontariff barriers</td>
<td>102</td>
</tr>
<tr>
<td>1986–94</td>
<td>Uruguay round</td>
<td>Tariffs, nontariff barriers, services, intellectual property, dispute settlement, textiles, agriculture, creation of WTO</td>
<td>123</td>
</tr>
<tr>
<td>2001–</td>
<td>Doha round</td>
<td>Agriculture, services, intellectual property, competition, investment, environment, dispute settlement</td>
<td>147</td>
</tr>
</tbody>
</table>

The sluggish pace of GATT negotiations led to an old joke that GATT really stood for Gentleman’s Agreement to Talk and Talk. The slow pace of international trade talks, however, is understandable, even sensible. Having dozens of nations agree to any treaty is a lengthy process. GATT often set up separate trading rules for certain industries, like agriculture, and separate trading rules for certain countries, like the low-income countries. There were rules, exceptions to rules, opportunities to opt out of rules, and precise wording to be fought over in every case.

### Watch It

Watch this video to review some of these major trade agreements.

Watch this video online: [https://youtu.be/27J3CByXKow](https://youtu.be/27J3CByXKow)

### Try It
TRADE POLICY: ORGANIZATIONS AND AGREEMENTS

Learning Objectives

- Explain the origin and role of the World Trade Organization (WTO)
- Discuss the significance and provide examples of regional trading agreements
- Analyze trade policy and evaluate long-term trends in barriers to trade

How Trade Policy Is Enacted: Globally, Regionally, and Nationally

Nations participate in global and regional trade agreements. They also develop their own national trade policies. The purpose of these agreements is to define what constitutes fair trading practices in different contexts.

The World Trade Organization

The World Trade Organization (WTO) was established in 1995, as the successor to the General Agreement on Tariffs and Trade (GATT), which was discussed in the last section. The WTO is committed to lowering barriers to trade. The world’s nations meet through the WTO to negotiate how they can reduce barriers to trade, such as tariffs. WTO negotiations happen in “rounds,” where all countries negotiate one agreement to encourage trade, take a year or two off, and then start negotiating a new agreement. The current round of negotiations is called the Doha Round because it was officially launched in Doha, the capital city of Qatar, in November 2001. In 2009, economists from the World Bank summarized recent research and found that the Doha round of negotiations would increase the size of the world economy by $160 billion to $385 billion per year, depending on the precise deal that ended up being negotiated.

In the context of a global economy that currently produces more than $30 trillion of goods and services each year, this amount is not huge: it is an increase of 1% or less. But before dismissing the gains from trade too quickly, it is worth remembering two points.
First, a gain of a few hundred billion dollars is enough money to deserve attention! Moreover, remember that this increase is not a one-time event; it would persist each year into the future.

Second, the estimate of gains may be on the low side because some of the gains from trade are not measured especially well in economic statistics. For example, it is difficult to measure the potential advantages to consumers of having a variety of products available and a greater degree of competition among producers. Perhaps the most important unmeasured factor is that trade between countries, especially when firms are splitting up the value chain of production, often involves a transfer of knowledge that can involve skills in production, technology, management, finance, and law.

Low-income countries benefit more from trade than high-income countries do. In some ways, the giant U.S. economy has less need for international trade, because it can already take advantage of internal trade within its economy. However, many smaller national economies around the world, in regions like Latin America, Africa, the Middle East, and Asia, have much more limited possibilities for trade inside their countries or their immediate regions. Without international trade, they may have little ability to benefit from comparative advantage, slicing up the value chain, or economies of scale. Moreover, smaller economies often have fewer competitive firms making goods within their economy, and thus firms have less pressure from other firms to provide the goods and prices that consumers want.

The economic gains from expanding international trade are measured in hundreds of billions of dollars, and the gains from international trade as a whole probably reach well into the trillions of dollars. The potential for gains from trade may be especially high among the smaller and lower-income countries of the world.

Like the GATT before it, the WTO is not a world government, with power to impose its decisions on others. The total staff of the WTO in 2013 is 629 people and its annual budget (as of 2012) is $196 million, which makes it smaller in size than many large universities.

Regional Trading Agreements

There are different types of economic integration across the globe, ranging from free trade agreements, in which participants allow each other’s imports without tariffs or quotas, to common markets, in which participants have a common external trade policy as well as free trade within the group, to full economic unions, in which, in addition to a common market, monetary and fiscal policies are coordinated. Many nations belong both to the World Trade Organization and to regional trading agreements.

The best known of these regional trading agreements is the European Union. In the years after World War II, leaders of several European nations reasoned that if they could tie their economies together more closely, they might be more likely to avoid another devastating war. Their efforts began with a free trade association, evolved into a common market, and then transformed into what is nearly a full economic union, known as the European Union. (The EU, as it is often called, has not included a common fiscal policy.) The EU has a number of goals. For example, in the early 2000s it introduced a common currency for Europe, the euro, and phased out most of the former national forms of money like the German mark and the French franc, though a few have retained their own currency. Another key element of the union is to eliminate barriers to the mobility of goods, labor, and capital across Europe.

For the United States, perhaps the best-known regional trading agreement is the North American Free Trade Agreement (NAFTA). The United States also participates in some less-prominent regional trading agreements, like the Caribbean Basin Initiative, which offers reduced tariffs for imports from these countries, and a free trade agreement with Israel.

The world has seen a flood of regional trading agreements in recent years. About 100 such agreements are now in place. A few of the more prominent ones are listed in Table 2. Some are just agreements to continue talking; others set specific goals for reducing tariffs, import quotas, and nontariff barriers. One economist described the current trade treaties as a “spaghetti bowl,” which is what a map with lines connecting all the countries with trade treaties looks like.

There is concern among economists who favor free trade that some of these regional agreements may promise free trade, but actually act as a way for the countries within the regional agreement to try to limit trade from anywhere else.
In some cases, the regional trade agreements may even conflict with the broader agreements of the World Trade Organization.

<table>
<thead>
<tr>
<th>Trade Agreements</th>
<th>Participating Countries</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asia Pacific Economic Cooperation (APEC)</td>
<td>Australia, Brunei, Canada, Chile, People’s Republic of China, Hong Kong, China, Indonesia, Japan, Republic of Korea, Malaysia, Mexico, New Zealand, Papua New Guinea, Peru, Philippines, Russia, Singapore, Chinese Taipei, Thailand, United States, Vietnam</td>
</tr>
<tr>
<td>European Union (EU)</td>
<td>Austria, Belgium, Bulgaria, Cyprus, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Ireland, Italy, Latvia, Lithuania, Luxembourg, Malta, Netherlands, Poland, Portugal, Romania, Slovakia, Slovenia, Spain, Sweden, United Kingdom</td>
</tr>
<tr>
<td>North America Free Trade Agreement (NAFTA)</td>
<td>Canada, Mexico, United States</td>
</tr>
<tr>
<td>Latin American Integration Association (LAIA)</td>
<td>Argentina, Bolivia, Brazil, Chile, Columbia, Ecuador, Mexico, Paraguay, Peru, Uruguay, Venezuela</td>
</tr>
<tr>
<td>Association of Southeast Asian Nations (ASEAN)</td>
<td>Brunei, Cambodia, Indonesia, Laos, Malaysia, Myanmar, Philippines, Singapore, Thailand, Vietnam</td>
</tr>
<tr>
<td>Southern African Development Community (SADC)</td>
<td>Angola, Botswana, Congo, Lesotho, Malawi, Mauritius, Mozambique, Namibia, Seychelles, South Africa, Swaziland, Tanzania, Zambia, Zimbabwe</td>
</tr>
</tbody>
</table>

National Trade Policies

Yet another dimension of trade policy, along with international and regional trade agreements, happens at the national level. Ideally, these policies do not conflict with the rules of the WTO and regional trade agreements. When there is an apparent conflict, the parent organization must adjudicate it. The United States, for example, imposes import quotas on sugar, because of a fear that such imports would drive down the price of sugar and thus injure domestic sugar producers. Why is sugar favored, while other products are not? Sometimes a product is protected because of historical practice. Sometimes it’s because a product has a particularly strong lobby. Recall, though, that trade barriers always end up costing a nation more than the benefits received by the protected group.

One of the jobs of the United States Department of Commerce is to determine if imports from other countries are being traded fairly. A common complaint is dumping, which means that foreign imports are being sold at less than their fair market value, i.e. their cost. The Commerce Department estimates a dumping “margin,” that is, the difference between price and cost. If Commerce determines that the import price is less than cost, they find that dumping has occurred. The United States International Trade Commission—another government agency—determines whether domestic industries have been substantially injured by the dumping, and if so, the President can impose tariffs in the amount of the dumping margin to offset the unfairly low price.
In the arena of trade policy, the battle often seems to be between national laws that increase protectionism and international agreements that try to reduce protectionism, like the WTO. Why would a country pass laws or negotiate agreements to shut out certain foreign products, like sugar or textiles, while simultaneously negotiating to reduce trade barriers in general? One plausible answer is that international trade agreements offer a method for countries to restrain their own special interests. A member of Congress can say to an industry lobbying for tariffs or quotas on imports: “Sure would like to help you, but that pesky WTO agreement just won’t let me.”

**Try It**
Visit this page in your course online to view this presentation.

**Watch It**
This video provides more information about trade blocs and the ways in which nations make arrangements regarding trade.
Watch this video online: https://youtu.be/YDUq0DINhYk
Watch this next video for a recent news example of an actual trade agreement between the United States and South Korea.
Watch this video online: https://youtu.be/EAh_eSbGKdl

**Long-Term Trends in Barriers to Trade**

In newspaper headlines, trade policy appears mostly as disputes and acrimony. Countries are almost constantly threatening to challenge the “unfair” trading practices of other nations. Cases are brought to the dispute settlement procedures of the WTO, the European Union, NAFTA, and other regional trading agreements. Politicians in national legislatures, goaded on by lobbyists, often threaten to pass bills that will “establish a fair playing field” or “prevent unfair trade”—although most such bills seek to accomplish these high-sounding goals by placing more restrictions on trade. Protesters in the streets may object to specific trade rules or to the entire practice of international trade.

Through all the controversy, the general trend for most of the last 60 years is clearly toward lower barriers to trade. The average level of tariffs on imported products charged by industrialized countries was 40% in 1946. By 1990, after decades of GATT negotiations, it was down to less than 5%. Indeed, one of the reasons that GATT negotiations shifted from focusing on tariff reduction in the early rounds to a broader agenda was that tariffs had been reduced so dramatically there was not much more to do in that area. U.S. tariffs have followed this general pattern: After rising sharply during the Great Depression, tariffs dropped off to less than 2% by the end of the century. Although measures of import quotas and nontariff barriers are less exact than those for tariffs, they generally appear to be at lower levels, too.

Thus, the last half-century has seen both a dramatic reduction in government-created barriers to trade, such as tariffs, import quotas, and nontariff barriers, and also a number of technological developments that have made international trade easier, like advances in transportation, communication, and information management. The result has been the powerful surge of international trade.

**Try It**
Visit this page in your course online to view this presentation.

**Glossary**

**common market**: economic agreement between countries to allow free trade in goods, services, labor, and financial capital between members while having a common external trade policy

**dumping**: selling imports at a price below fair market value, i.e. cost
economic union: economic agreement between countries to allow free trade between members, a common external trade policy, and coordinated monetary and fiscal policies

free trade agreement: economic agreement between countries to allow free trade between members

PUTTING IT TOGETHER: GLOBALIZATION AND INTERNATIONAL TRADE

The goal of this module was to analyze the benefits and costs of international trade, and to determine the extent to which barriers to international trade are warranted.

You learned how to:

- Define and calculate comparative and absolute advantage
- Define and calculate gains from trade
- Understand the way imports and exports impact different actors in the economy (businesses, consumers, and workers)
- Explain how globalization has increased over time, especially over the last several decades
- Understand the way barriers to trade (e.g. tariffs, quotas and non-tariff barriers) affect business, consumers and workers in the economy.
- Differentiate between alternative international trade regimes and how they impact global trade

You learned that trade based on comparative advantage will maximize an individual’s or a nation’s income, but that there will be winners and losers to trade. For example, employees and owners of a firm that loses business to foreign imports are worse off, even though their loss is less than the gain to consumers. The challenge for policymakers is how to compensate the losers while capturing the gains from trade. Similarly, protectionism benefits some workers and businesses at the expense of other workers and businesses and at the expense of consumers. In this case, the losses to the latter groups are larger than the gains to the former groups. In that sense, protectionism makes a country worse off.
WHY IT MATTERS: EXCHANGE RATES AND INTERNATIONAL FINANCE

Why learn about exchange rates and the way they influence international trade?

Is a Stronger Dollar Good for the U.S. Economy?

From 2002 to 2008, the U.S. dollar lost more than a quarter of its value in foreign currency markets. On January 1, 2002, one dollar was worth 1.11 euros. On April 24, 2008 it hit its lowest point with a dollar being worth 0.64 euros. During this period, the trade deficit between the United States and the European Union grew from a yearly total of approximately –85.7 billion dollars in 2002 to 95.8 billion dollars in 2008. Was this a good thing or a bad thing for the U.S. economy?
We live in a global world. U.S. consumers buy trillions of dollars worth of imported goods and services each year, not just from the European Union, but from all over the world. U.S. businesses sell trillions of dollars’ worth of exports. U.S. citizens, businesses, and governments invest trillions of dollars abroad every year. Foreign investors, businesses, and governments invest trillions of dollars in the United States each year. Indeed, foreigners are a major buyer of U.S. federal debt.

Many people feel that a weaker dollar is bad for America, that it’s an indication of a weak economy. But is it? This section will help answer that question.

The world has over 150 different currencies, from the Afghanistan afghani and the Albanian lek all the way through the alphabet to the Zambian kwacha and the Zimbabwean dollar. For international economic transactions, households or firms will wish to exchange one currency for another. Perhaps the need for exchanging currencies will come from a German firm that exports products to Russia, but then wishes to exchange the Russian rubles it has earned for euros, so that the firm can pay its workers and suppliers in Germany. Perhaps it will be a South African firm that wishes to purchase a mining operation in Angola, but to make the purchase it must convert South African rand to Angolan kwanza. Perhaps it will be an American tourist visiting China, who wishes to convert U.S. dollars to Chinese yuan to pay the hotel bill.

Exchange rates can sometimes change very swiftly. For example, in the United Kingdom the pound was worth about $1.50 just before the nation voted to leave the European Union (also known as the Brexit vote), but fell to $1.37 just after the vote and continued falling to reach 30-year lows a few months later. For firms engaged in international buying, selling, lending, and borrowing, these swings in exchange rates can have an enormous effect on profits.

This module discusses the international dimension of money, which involves conversions from one currency to another at an exchange rate. An exchange rate is nothing more than a price—that is, the price of one currency in terms of another currency—and so we can analyze it with the tools of supply and demand. First, we’ll learn about foreign exchange markets: their size, their main participants, and the vocabulary for discussing movements of exchange rates. Next, we’ll use demand and supply graphs to analyze some of the main factors that cause shifts in exchange rates. Finally, we’ll bring the central bank and monetary policy back into the picture. Each country must decide whether to allow the market to determine its exchange rate, or have the central bank intervene. All the choices for exchange rate policy involve distinctive tradeoffs and risks.

INTRODUCTION TO THE FOREIGN EXCHANGE MARKET

What you’ll learn to do: define currency exchange rates and explain how they influence trade balances
In the foreign exchange market, people and firms exchange one currency to purchase another currency. This market is influenced by both demand and supply:

- The demand for dollars comes from those U.S. export firms seeking to convert their earnings in foreign currency back into U.S. dollars; foreign tourists converting their earnings in a foreign currency back into U.S. dollars; and foreign investors seeking to make financial investments in the U.S. economy.
- On the supply side of the foreign exchange market for the trading of U.S. dollars are foreign firms that have sold imports in the U.S. economy and are seeking to convert their earnings back to their home currency; U.S. tourists abroad; and U.S. investors seeking to make financial investments in foreign economies.

A stronger currency benefits those who are buying with that currency and injures those who are selling. In this section, you’ll learn about why this is.

---

**THE FOREIGN EXCHANGE MARKET**

**Learning Objectives**

- Explain the foreign exchange market and the main groups of people or firms who participate in the market
- Describe different types of investments like foreign direct investment (FDI), portfolio investment, and hedging
Most countries have their own currencies, but not all. Sometimes small economies use the currency of an economically larger neighbor. For example, Ecuador, El Salvador, and Panama have decided to dollarize—that is, to use the U.S. dollar as their currency. Sometimes nations share a common currency. The best example of a common currency is the Euro, a common currency used by 19 members of the European Union. With these exceptions duly noted, most international transactions require participants to convert from one currency to another when selling, buying, hiring, borrowing, traveling, or investing across national borders. The market in which people or firms use one currency to purchase another currency is called the foreign exchange market.

Try It

Visit this page in your course online to check your understanding.

Every exchange rate is a price—the price of one currency expressed in terms of units of another currency. The key framework for analyzing prices, whether in this course, any other economics course, in public policy, or business examples, is supply and demand in markets.

Watch It

Watch this video for a brief introduction to the foreign exchange market (or forex).
Watch this video online: https://youtu.be/-qvrRRTBYAk

LINK IT UP

Visit this website for an exchange rate calculator.

The Extraordinary Size of the Foreign Exchange Markets

The quantities traded in foreign exchange markets are breathtaking. A survey done in April, 2013 by the Bank of International Settlements, an international organization for banks and the financial industry, found that $5.3 trillion per day was traded on foreign exchange markets, which makes the foreign exchange market the largest market in the world economy. In contrast, 2013 U.S. real GDP was $15.8 trillion per year.

Table 1 shows the currencies most commonly traded on foreign exchange markets. The foreign exchange market is dominated by the U.S. dollar, the Euro, the Japanese yen, and the British pound.

<table>
<thead>
<tr>
<th>Currency</th>
<th>% Daily Share</th>
</tr>
</thead>
<tbody>
<tr>
<td>U.S. dollar</td>
<td>87.6%</td>
</tr>
<tr>
<td>Euro</td>
<td>31.3%</td>
</tr>
<tr>
<td>Japanese yen</td>
<td>21.6%</td>
</tr>
<tr>
<td>British pound</td>
<td>12.8%</td>
</tr>
<tr>
<td>Australian dollar</td>
<td>6.9%</td>
</tr>
<tr>
<td>Canadian dollar</td>
<td>5.1%</td>
</tr>
<tr>
<td>Swiss franc</td>
<td>4.8%</td>
</tr>
<tr>
<td>Chinese yuan</td>
<td>2.6%</td>
</tr>
</tbody>
</table>
Demanders and Suppliers of Currency in Foreign Exchange Markets

In foreign exchange markets, demand and supply become closely interrelated, because a person or firm who demands one currency must at the same time supply another currency—and vice versa. To get a sense of this, it is useful to consider four groups of people or firms who participate in the market: (1) firms that import or export goods and services; (2) tourists visiting other countries; (3) international investors buying ownership (or part-ownership) in a foreign firm; (4) international investors making financial investments that do not involve ownership. Let's consider these categories in turn.

Firms that sell exports or buy imports find that their costs for workers, suppliers, and investors are measured in the currency of the nation where their production occurs, but their revenues from sales are measured in the currency of the different nation where their sales happened. So, a Chinese firm exporting abroad will earn some other currency—say, U.S. dollars—but will need Chinese yuan to pay the workers, suppliers, and investors who are based in China. In the foreign exchange markets, this firm will be a supplier of U.S. dollars and a demander of Chinese yuan.

International tourists need foreign currency for expenses in the country they are visiting; they will supply their home currency to receive the foreign currency. For example, an American tourist who is visiting China will supply U.S. dollars into the foreign exchange market and demand Chinese yuan.

Financial investments that cross international boundaries, and require exchanging currency, are often divided into two categories. **Foreign direct investment (FDI)** refers to purchasing (at least ten percent) ownership in a firm in another country or starting up a new enterprise in a foreign country. For example, in 2008 the Belgian beer-brewing company InBev bought the U.S. beer-maker Anheuser-Busch for $52 billion. To make this purchase of a U.S. firm, InBev had to supply euros (the currency of Belgium) to the foreign exchange market and demand U.S. dollars.

The other kind of international financial investment, **portfolio investment**, involves a purely financial investment that does not entail any management responsibility. An example would be a U.S. financial investor who purchased bonds issued by the government of the United Kingdom, or deposited money in a British bank. To make such investments, the American investor would supply U.S. dollars in the foreign exchange market and demand British pounds.

Portfolio investment is often linked to expectations about how exchange rates will shift. Look at a U.S. financial investor who is considering purchasing bonds issued in the United Kingdom. For simplicity, ignore any interest paid by the bond (which will be small in the short run anyway) and focus on exchange rates. Say that a British pound is currently worth $1.50 in U.S. currency. However, the investor believes that in a month, the British pound will be worth $1.60 in U.S. currency. Thus, as Figure 2(a) shows, this investor would change $24,000 for 16,000 British pounds. In a month, if the pound is indeed worth $1.60, then the portfolio investor can trade back to U.S. dollars at the new exchange rate, and have $25,600—a nice profit. A portfolio investor who believes that the foreign exchange rate for the pound will work in the opposite direction can also invest accordingly. Say that an investor expects that the pound, now worth $1.50 in U.S. currency, will decline to $1.40. Then, as shown in Figure 2(b), that investor could start off with £20,000 in British currency (borrowing the money if necessary), convert it to $30,000 in U.S. currency, wait a month, and then convert back to approximately £21,429 in British currency—again making a nice profit. Of course, this kind of investing comes without guarantees, and an investor may suffer losses if the exchange rates do not move as predicted.
Figure 1. A Portfolio Investor Trying to Benefit from Exchange Rate Movements. Expectations of the future value of a currency can drive demand and supply of that currency in foreign exchange markets.

Many portfolio investment decisions are not as simple as betting that the value of the currency will change in one direction or the other. Instead, they involve firms trying to protect themselves from movements in exchange rates. Imagine you are running a U.S. firm that is exporting to France. You have signed a contract to deliver certain products and will receive 1 million euros a year from now. But you do not know how much this contract will be worth in U.S. dollars, because the dollar/euro exchange rate can fluctuate in the next year. Let’s say you want to know for sure what the contract will be worth, and not take a risk that the euro will be worth less in U.S. dollars than it currently is. You can hedge, which means using a financial transaction to protect yourself against currency risk. Specifically, you can sign a financial contract and pay a fee that guarantees you a certain exchange rate one year from now—regardless of what the market exchange rate is at that time. Now, it is possible that the euro will be worth more in dollars a year from now, so your hedging contract will be unnecessary, and you will have paid a fee for nothing. But if the value of the euro in dollars declines, then you are protected by the hedge. Financial contracts like hedging, where parties wish to be protected against exchange rate movements, also commonly lead to a series of portfolio investments by the firm that is receiving a fee to provide the hedge.

Both foreign direct investment and portfolio investment involve an investor who supplies domestic currency and demands a foreign currency. With portfolio investment less than ten percent of a company is purchased. As such, portfolio investment is often made with a short term focus. With foreign direct investment more than ten percent of a company is purchased and the investor typically assumes some managerial responsibility; thus foreign direct investment tends to have a more long-run focus. As a practical matter, portfolio investments can be withdrawn from a country much more quickly than foreign direct investments. A U.S. portfolio investor who wants to buy or sell bonds issued by the government of the United Kingdom can do so with a phone call or a few clicks of a computer key. However, a U.S. firm that wants to buy or sell a company, such as one that manufactures automobile parts in the United Kingdom, will find that planning and carrying out the transaction takes a few weeks, even months. Table 2 summarizes the main categories of demanders and suppliers of currency.

Table 2. The Demand and Supply Line-ups in Foreign Exchange Markets

<table>
<thead>
<tr>
<th>Demand for the U.S. Dollar Comes from...</th>
<th>Supply of the U.S. Dollar Comes from...</th>
</tr>
</thead>
<tbody>
<tr>
<td>A U.S. exporting firm that earned foreign currency and is trying to pay U.S.-based expenses</td>
<td>A foreign firm that has sold imported goods in the United States, earned U.S. dollars, and is trying to pay expenses incurred in its home country</td>
</tr>
<tr>
<td>Foreign tourists visiting the United States</td>
<td>U.S. tourists leaving to visit other countries</td>
</tr>
</tbody>
</table>
Demand for the U.S. Dollar Comes from...
- Foreign investors who wish to make direct investments in the U.S. economy
- Foreign investors who wish to make portfolio investments in the U.S. economy

Supply of the U.S. Dollar Comes from...
- U.S. investors who want to make foreign direct investments in other countries
- U.S. investors who want to make portfolio investments in other countries

Try It
Visit this page in your course online to check your understanding.

Participants in the Exchange Rate Market

The foreign exchange market does not involve the ultimate suppliers and demanders of foreign exchange literally seeking each other out. If Martina decides to leave her home in Venezuela and take a trip in the United States, she does not need to find a U.S. citizen who is planning to take a vacation in Venezuela and arrange a person-to-person currency trade. Instead, the foreign exchange market works through financial institutions, and it operates on several levels.

Most people and firms who are exchanging a substantial quantity of currency go to a bank, and most banks provide foreign exchange as a service to customers. These banks (and a few other firms), known as dealers, then trade the foreign exchange. This is called the interbank market.

In the world economy, roughly 2,000 firms are foreign exchange dealers. The U.S. economy has less than 100 foreign exchange dealers, but the largest 12 or so dealers carry out more than half the total transactions. The foreign exchange market has no central location, but the major dealers keep a close watch on each other at all times.

The foreign exchange market is huge not because of the demands of tourists, firms, or even foreign direct investment, but instead because of portfolio investment and the actions of interlocking foreign exchange dealers. International tourism is a very large industry, involving about $1 trillion per year. Global exports are about 23% of global GDP; which is about $18 trillion per year. Foreign direct investment totaled about $1.4 trillion in 2012. These quantities are dwarfed, however, by the $5.3 trillion per day being traded in foreign exchange markets. Most transactions in the foreign exchange market are for portfolio investment—relatively short-term movements of financial capital between currencies—and because of the actions of the large foreign exchange dealers as they constantly buy and sell with each other.

Watch It
This video clip will give you an example of how money is exchanged on the foreign exchange market and how that affects international trade. We’ll dig into some of the concepts in further detail in the readings that follow.
Visit this page in your course online to view this presentation.

Glossary
- **dollarize**: a country that is not the United States uses the U.S. dollar as its currency
- **exchange rate**: the price of one currency expressed in terms of units of another currency
- **foreign direct investment (FDI)**: purchasing more than ten percent of a firm or starting a new enterprise in another country
- **foreign exchange market**: the market in which people use one currency to buy another currency
- **hedge**: using a financial transaction as protection against risk
**STRENGTHENING AND WEAKENING CURRENCY**

**Learning Objectives**

- Describe the exchange rate when a currency increases in value and a currency decreases in value
- Identify who benefits from a stronger currency and benefits from a weaker currency

**Strengthening and Weakening Currency**

When the value of a currency rises, so that the currency exchanges for more of other currencies, the exchange rate is described as *appreciating* or “strengthening.” When the value of a currency falls, so that a currency trades for less of other currencies, the exchange rate is described as *depreciating* or “weakening.”

To illustrate the use of these terms, consider the exchange rate between the U.S. dollar and the Canadian dollar since 1980, in Figure 1(a). The vertical axis in Figure 1(a) shows the price of $1 in U.S. currency, measured in terms of Canadian currency. Clearly, exchange rates can move up and down substantially. A U.S. dollar traded for $1.17 Canadian in 1980. The U.S. dollar appreciated or strengthened to $1.39 Canadian in 1986, depreciated or weakened to $1.15 Canadian in 1991, and then appreciated or strengthened to $1.60 Canadian by early in 2002, fell to roughly $1.20 Canadian in 2009, and then had a sharp spike up and decline in 2009 and 2010. In May of 2017, the U.S. dollar stood at $1.36 Canadian. The units in which we measure exchange rates can be confusing, because we measure the exchange rate of the U.S. dollar exchange using a different currency—the Canadian dollar. However, exchange rates always measure the price (or value) of one unit of currency by using a different currency.
Figure 1. Exchange rates tend to fluctuate substantially, even between bordering countries such as the United States and Canada. By looking closely at the time values (the years vary slightly on these graphs), it is clear that the values in part (a) are a mirror image of part (b), which demonstrates that the depreciation of one currency correlates to the appreciation of the other and vice versa. This means that when comparing the exchange rates between two countries (in this case, the United States and Canada), the depreciation (or weakening) of one currency (the U.S. dollar for this example) indicates the appreciation (or strengthening) of the other currency (which in this example is the Canadian dollar). (Source: Federal Reserve Economic Data (FRED) https://research.stlouisfed.org/fred2/series/EXCAUS)

In looking at the exchange rate between two currencies, the appreciation or strengthening of one currency must mean the depreciation or weakening of the other. Figure 1(b) shows the exchange rate for the Canadian dollar, measured in terms of U.S. dollars. The exchange rate of the U.S. dollar measured in Canadian dollars, shown in Figure 1(a), is a perfect mirror image with the exchange rate of the Canadian dollar measured in U.S. dollars, shown in Figure 15(b). A fall in the Canada $/U.S. $ ratio means a rise in the U.S. $/Canada $ ratio, and vice versa.

Try It
Visit this page in your course online to check your understanding.

Watch It
This video explains how the exchange rate is determined using supply and demand.
With the price of a typical good or service, it is clear that higher prices benefit sellers and hurt buyers, while lower prices benefit buyers and hurt sellers. In the case of exchange rates, where the buyers and sellers are not always intuitively obvious, it is useful to trace through how different participants in the market will be affected by a stronger or weaker currency. Consider, for example, the impact of a stronger U.S. dollar on six different groups of economic actors, as shown in Figure 2: (1) U.S. exporters selling abroad; (2) foreign exporters (that is, firms selling imports in the U.S. economy); (3) U.S. tourists abroad; (4) foreign tourists visiting the United States; (5) U.S. investors (either foreign direct investment or portfolio investment) considering opportunities in other countries; (6) and foreign investors considering opportunities in the U.S. economy.

**Figure 2. How Do Exchange Rate Movements Affect Each Group?** Exchange rate movements affect exporters, tourists, and international investors in different ways.

For a U.S. firm selling abroad, a stronger U.S. dollar is a curse. A strong U.S. dollar means that foreign currencies are correspondingly weak. When this exporting firm earns foreign currencies through its export sales, and then converts them back to U.S. dollars to pay workers, suppliers, and investors, the stronger dollar means that the foreign currency buys fewer U.S. dollars than if the currency had not strengthened, and that the firm’s profits (as measured in dollars) fall. As a result, the firm may choose to reduce its exports, or it may raise its selling price, which will also tend to reduce its exports. In this way, a stronger currency reduces a country’s exports.

Conversely, for a foreign firm selling in the U.S. economy, a stronger dollar is a blessing. Each dollar earned through export sales, when traded back into the home currency of the exporting firm, will now buy more of the home currency than expected before the dollar had strengthened. As a result, the stronger dollar means that the importing firm will earn higher profits than expected. The firm will seek to expand its sales in the U.S. economy, or it may reduce prices, which will also lead to expanded sales. In this way, a stronger U.S. dollar means that consumers will purchase more from foreign producers, expanding the country’s level of imports.

For a U.S. tourist abroad, who is exchanging U.S. dollars for foreign currency as necessary, a stronger U.S. dollar is a benefit. The tourist receives more foreign currency for each U.S. dollar, and consequently the cost of the trip in U.S. dollars is lower. When a country’s currency is strong, it is a good time for citizens of that country to tour abroad. Imagine a U.S. tourist who has saved up $5,000 for a trip to South Africa. In January 2008, $1 bought 7 South African rand, so the tourist had 35,000 rand to spend. In January 2009, $1 bought 10 rand, so the tourist had 50,000 rand to spend. By January 2010, $1 bought only 7.5 rand. Clearly, 2009 was the year for U.S. tourists to visit South Africa. For foreign visitors to the United States, the opposite pattern holds true. A relatively stronger U.S. dollar means that their own currencies are relatively weaker, so that as they shift from their own currency to U.S. dollars, they have fewer U.S. dollars than previously. When a country’s currency is strong, it is not an especially good time for foreign tourists to visit.

A stronger dollar injures the prospects of a U.S. financial investor who has already invested money in another country. A U.S. financial investor abroad must first convert U.S. dollars to a foreign currency, invest in a foreign country, and then later convert that foreign currency back to U.S. dollars. If in the meantime the U.S. dollar becomes stronger and the
foreign currency becomes weaker, then when the investor converts back to U.S. dollars, the rate of return on that investment will be less than originally expected at the time it was made.

However, a stronger U.S. dollar boosts the returns of a foreign investor putting money into a U.S. investment. That foreign investor converts from the home currency to U.S. dollars and seeks a U.S. investment, while later planning to switch back to the home currency. If, in the meantime, the dollar grows stronger, then when the time comes to convert from U.S. dollars back to the foreign currency, the investor will receive more foreign currency than expected at the time the original investment was made.

The preceding paragraphs all focus on the case where the U.S. dollar becomes stronger. The corresponding happy or unhappy economic reactions are illustrated in the first column of Figure 2. The following feature centers the analysis on the opposite: a weaker dollar.

**EFFECTS OF A WEAKER DOLLAR**

Let's examine the effects of a weaker dollar in various scenarios.

**Scenario 1: What will happen to the price of a Ford pickup truck in the U.K. if the value of the dollar weakens?**
First, we note that the demand for U.S. exports is a function of the price of those exports, which depends on the dollar price of those goods and the exchange rate of the dollar in terms of foreign currency. For example, a Ford pickup truck costs $25,000 in the United States. When it is sold in the United Kingdom, the price is $25,000/$1.50 per British pound, or £16,667. The dollar affects the price faced by foreigners who may purchase U.S. exports.

Next, consider that, if the dollar weakens, the pound rises in value. If the pound rises to $2.00 per pound, then the price of a Ford pickup is now $25,000/$2.00 = £12,500. A weaker dollar means the foreign currency buys more dollars, which means that U.S. exports appear less expensive.

From this, we conclude that a weaker U.S. dollar leads to an increase in U.S. exports. For a foreign exporter, the outcome is just the opposite.

**Scenario 2: How does a depreciating dollar affect the exchange of British beer in the United States?**
Suppose a brewery in England is interested in selling its Bass Ale to a grocery store in the United States. If the price of a six pack of Bass Ale is £6.00 and the exchange rate is $1.50 per British pound, the price for the grocery store is 6.00 × $1.50 = $9.00 per six pack. If the dollar weakens to $2.00 per pound, the price of Bass Ale is now 6.00 × $2.00 = $12.

We can conclude that from the perspective of U.S. purchasers, a weaker dollar means that foreign currency is more expensive, which means that foreign goods are more expensive also. This leads to a decrease in U.S. imports, which is bad for the foreign exporter.

**Scenario 3: How does a weaker dollar affect U.S. tourists traveling abroad?**
U.S. tourists going abroad face the same situation as a U.S. importer—they are purchasing a foreign trip. A weaker dollar means that their trip will cost more, since a given expenditure of foreign currency (e.g., hotel bill) will take more dollars. The result is that the tourist may not stay as long abroad, and some may choose not to travel at all.

**Scenario 4: How does a weaker dollar affect foreign tourists traveling to the United States?**
Consider that, for the foreign tourist to the United States, a weaker dollar is a boon. It means their currency goes further, so the cost of a trip to the United States will be less. Foreigners may choose to take longer trips to the United States, and more foreign tourists may decide to take U.S. trips.

**Scenario 5: How does a weaker dollar affect investments?**
A U.S. investor abroad faces the same situation as a U.S. importer—they are purchasing a foreign asset. A U.S. investor will see a weaker dollar as an increase in the “price” of investment, since the same number of dollars will buy less foreign currency and thus less foreign assets. This should decrease the amount of U.S. investment abroad.

Foreign investors in the United States will have the opposite experience. Since foreign currency buys more dollars, they will likely invest in more U.S. assets.

At this point, you should have a good sense of the major players in the foreign exchange market: firms involved in international trade, tourists, international financial investors, banks, and foreign exchange dealers. The next section demonstrates in more detail how the tools of demand and supply can be used in foreign exchange markets to explain the underlying causes of stronger and weaker currencies.
WHY IS A STRONGER CURRENCY NOT NECESSARILY BETTER?

One common misunderstanding about exchange rates is that a “stronger” or “appreciating” currency must be better than a “weaker” or “depreciating” currency. After all, is it not obvious that “strong” is better than “weak”? But do not let the terminology confuse you. When a currency becomes stronger, so that it purchases more of other currencies, it benefits some in the economy and injures others. Stronger currency is not necessarily better, it is just different.

Try It

Visit this page in your course online to check your understanding.

Watch It

Watch this video to practice analyzing the impact of shifters on supply and demand in the foreign exchange market. Watch this video online: https://youtu.be/hmbs_06LnS8

Try It

These questions allow you to get as much practice as you need, as you can click the link at the top of the first question (“Try another version of these questions”) to get a new set of questions. Practice until you feel comfortable doing the questions. Visit this page in your course online to practice before taking the quiz.

Glossary

**appreciating**: when a currency is worth more in terms of other currencies; also called “strengthening”

**depreciating**: when a currency is worth less in terms of other currencies; also called “weakening”

INTRODUCTION TO EXCHANGE RATES AND PURCHASING POWER
What you'll learn to do: analyze how supply and demand affects currencies and exchange rates

Have you ever considered traveling abroad to a country where you can get more bank for your buck? Maybe you could stock up on clothes, movies, or just enjoy paying less for food? Why do you think that happens? In this section, you'll learn about how variations in supply and demand between foreign currencies affect the purchasing power of your money.

DEMAND AND SUPPLY SHIFTS IN FOREIGN EXCHANGE MARKETS

Learning Objectives

- Explain the factors that cause the demand and supply of foreign currencies to shift
- Define arbitrage and the importance of purchasing power parity

Demand and Supply Shifts in Foreign Exchange Markets

The foreign exchange market involves firms, households, and investors who demand and supply currencies coming together through their banks and the key foreign exchange dealers. Figure 1(a) offers an example for the exchange rate between the U.S. dollar and the Mexican peso. The vertical axis shows the exchange rate for U.S. dollars, which in this case is measured in pesos. The horizontal axis shows the quantity of U.S. dollars being traded in the foreign exchange market.
market each day. The demand curve (D) for U.S. dollars intersects with the supply curve (S) of U.S. dollars at the equilibrium point (E), which is an exchange rate of 10 pesos per dollar and a total volume of $8.5 billion.

Figure 1. Demand and Supply for the U.S. Dollar and Mexican Peso Exchange Rate. (a) The quantity measured on the horizontal axis is in U.S. dollars, and the exchange rate on the vertical axis is the price of U.S. dollars measured in Mexican pesos. (b) The quantity measured on the horizontal axis is in Mexican pesos, while the price on the vertical axis is the price of pesos measured in U.S. dollars. In both graphs, the equilibrium exchange rate occurs at point E, at the intersection of the demand curve (D) and the supply curve (S).

Figure 1(b) presents the same demand and supply information from the perspective of the Mexican peso. The vertical axis shows the exchange rate for Mexican pesos, which is measured in U.S. dollars. The horizontal axis shows the quantity of Mexican pesos traded in the foreign exchange market. The demand curve (D) for Mexican pesos intersects with the supply curve (S) of Mexican pesos at the equilibrium point (E), which is an exchange rate of 10 cents in U.S. currency for each Mexican peso and a total volume of 85 billion pesos. Note that the two exchange rates are inverses: 10 pesos per dollar is the same as 10 cents per peso (or $0.10 per peso). In the actual foreign exchange market, almost all of the trading for Mexican pesos is done for U.S. dollars. What factors would cause the demand or supply to shift, thus leading to a change in the equilibrium exchange rate? Read on to discover the answer to this question.

Try It

Visit this page in your course online to view this presentation.

Expectations about Future Exchange Rates

One reason to demand a currency on the foreign exchange market is the belief that the value of the currency is about to increase. One reason to supply a currency—that is, sell it on the foreign exchange market—is the expectation that the value of the currency is about to decline. For example, imagine that a leading business newspaper, like the *Wall Street Journal* or the *Financial Times*, runs an article predicting that the Mexican peso will appreciate in value. The likely effects of such an article are illustrated in the interactive graph below (Figure 2). Demand for the Mexican peso shifts to the right, from D₀ to D₁, as investors become eager to purchase pesos. Conversely, the supply of pesos shifts to the left, from S₀ to S₁, because investors will be less willing to give them up. The result is that the equilibrium exchange rate rises from 10 cents/peso to 12 cents/peso and the equilibrium exchange rate rises from 85 billion to 90 billion pesos as the equilibrium moves from E₀ to E₁.

Visit this page in your course online to view this presentation.

Figure 2 (Interactive Graph). Exchange Rate Market for Mexican Peso Reacts to Expectations about Future Exchange Rates.

Figure 2 also illustrates some peculiar traits of supply and demand diagrams in the foreign exchange market. In contrast to all the other cases of supply and demand you have considered, in the foreign exchange market, supply and demand...
typically both move at the same time. Groups of participants in the foreign exchange market like firms and investors include some who are buyers and some who are sellers. An expectation of a future shift in the exchange rate affects both buyers and sellers—that is, it affects both demand and supply for a currency.

The shifts in demand and supply curves both cause the exchange rate to shift in the same direction; in this example, they both make the peso exchange rate stronger. However, the shifts in demand and supply work in opposing directions on the quantity traded. In this example, the rising demand for pesos is causing the quantity to rise while the falling supply of pesos is causing quantity to fall. In this specific example, the result is a higher quantity. But in other cases, the result could be that quantity remains unchanged or declines.

This example also helps to explain why exchange rates often move quite substantially in a short period of a few weeks or months. When investors expect a country’s currency to strengthen in the future, they buy the currency and cause it to appreciate immediately. The appreciation of the currency can lead other investors to believe that future appreciation is likely—and thus lead to even further appreciation. Similarly, a fear that a currency might weaken quickly leads to an actual weakening of the currency, which often reinforces the belief that the currency is going to weaken further. Thus, beliefs about the future path of exchange rates can be self-reinforcing, at least for a time, and a large share of the trading in foreign exchange markets involves dealers trying to outguess each other on what direction exchange rates will move next.

Differences across Countries in Rates of Return

The motivation for investment, whether domestic or foreign, is to earn a return. If rates of return in a country look relatively high, then that country will tend to attract funds from abroad. Conversely, if rates of return in a country look relatively low, then funds will tend to flee to other economies. Changes in the expected rate of return will shift demand and supply for a currency. For example, imagine that interest rates rise in the United States as compared with Mexico. Thus, financial investments in the United States promise a higher return than they previously did. As a result, more investors will demand U.S. dollars so that they can buy interest-bearing assets and fewer investors will be willing to supply U.S. dollars to foreign exchange markets. Demand for the U.S. dollar will shift to the right, from $D_0$ to $D_1$, and supply will shift to the left, from $S_0$ to $S_1$, as shown in the interactive graph below (Figure 3). The new equilibrium ($E_1$), will occur at an exchange rate of nine pesos/dollar and the same quantity of $8.5 billion. Thus, a higher interest rate or rate of return relative to other countries leads a nation's currency to appreciate or strengthen, and a lower interest rate relative to other countries leads a nation's currency to depreciate or weaken. Since a nation's central bank can use monetary policy to affect its interest rates, a central bank can also cause changes in exchange rates.

Relative Inflation

If a country experiences a relatively high inflation rate compared with other economies, then the buying power of its currency is eroding, which will tend to discourage anyone from wanting to acquire or to hold the currency. Figure 4 (the interactive graph below) shows an example based on an actual episode concerning the Mexican peso. In 1986–87, Mexico experienced an inflation rate of over 200%. Not surprisingly, as inflation dramatically decreased the purchasing power of the peso in Mexico, the exchange rate value of the peso declined as well. As shown in Figure 4, demand for the peso on foreign exchange markets decreased from $D_0$ to $D_1$, while supply of the peso increased from $S_0$ to $S_1$. The equilibrium exchange rate fell from $2.50 per peso at the original equilibrium ($E_0$) to $0.50 per peso at the new equilibrium ($E_1$). In this example, the quantity of pesos traded on foreign exchange markets remained the same, even as the exchange rate shifted.

Purchasing Power Parity
Over the long term, exchange rates must bear some relationship to the buying power of the currency in terms of goods that are internationally traded. If at a certain exchange rate it was much cheaper to buy internationally traded goods—such as oil, steel, computers, and cars—in one country than in another country, businesses would start buying in the cheap country, selling in other countries, and pocketing the profits.

For example, if a U.S. dollar is worth $1.60 in Canadian currency, then a car that sells for $20,000 in the United States should sell for $32,000 in Canada. If the price of cars in Canada was much lower than $32,000, then at least some U.S. car-buyers would convert their U.S. dollars to Canadian dollars and buy their cars in Canada. If the price of cars was much higher than $32,000 in this example, then at least some Canadian buyers would convert their Canadian dollars to U.S. dollars and go to the United States to purchase their cars. This is known as arbitrage, the process of buying and selling goods or currencies across international borders at a profit. It may occur slowly, but over time, it will force prices and exchange rates to align so that the price of internationally traded goods is similar in all countries.

The exchange rate that equalizes the prices of internationally traded goods across countries is called the purchasing power parity (PPP) exchange rate. A group of economists at the International Comparison Program, run by the World Bank, have calculated the PPP exchange rate for all countries, based on detailed studies of the prices and quantities of internationally tradable goods.

Watch It

In this video, Alex shows you an example of purchasing power parity while on a trip to India.
Visit this page in your course online to view this presentation.

The purchasing power parity exchange rate has two functions. First, PPP exchange rates are often used for international comparison of GDP and other economic statistics. Imagine that you are preparing a table showing the size of GDP in many countries in several recent years, and for ease of comparison, you are converting all the values into U.S. dollars. When you insert the value for Japan, you need to use a yen/dollar exchange rate. But should you use the market exchange rate or the PPP exchange rate? Market exchange rates bounce around. In summer 2008, the exchange rate was 108 yen/dollar, but in late 2009 the U.S. dollar exchange rate versus the yen was 90 yen/dollar. For simplicity, say that Japan’s GDP was ¥500 trillion in both 2008 and 2009. If you use the market exchange rates, then Japan’s GDP will be $4.6 trillion in 2008 (that is, ¥500 trillion /¥108/dollar) and $5.5 trillion in 2009 (that is, ¥500 trillion /¥90/dollar).

Of course, it is not true that Japan’s economy increased enormously in 2009—in fact, Japan had a recession like much of the rest of the world. The misleading appearance of a booming Japanese economy occurs only because we used the market exchange rate, which often has short-run rises and falls. However, PPP exchange rates stay fairly constant and change only modestly, if at all, from year to year.

The second function of PPP is that exchanges rates will often get closer and closer to it as time passes. It is true that in the short run and medium run, as exchange rates adjust to relative inflation rates, rates of return, and to expectations about how interest rates and inflation will shift, the exchange rates will often move away from the PPP exchange rate for a time. But, knowing the PPP will allow you to track and predict exchange rate relationships.

Food For Thought

One interesting way to think about purchasing power parity is by comparing the price of a hamburger across different countries. Initially just for fun, The Economist began comparing the price of a BigMac in the BigMac Index between various countries in 1986, and continues to do so today as a simple way to see how currency may be either undervalued or overvalued.

Try It

Visit this page in your course online to view this presentation.
INTRODUCTION TO EXCHANGE RATES AND THE TRADE BALANCE

What you’ll learn to do: explain how the balance of trade (surplus or deficit) affects the domestic economy
In this section, you will learn how fluctuations in exchange rates affect imports and exports, and how changes in imports and exports affect the domestic economy.

MACROECONOMIC EFFECTS OF EXCHANGE RATES

Learning Objectives

- Explain how changes in exchange rates influence aggregate demand and supply

Exchange Rates, Aggregate Demand, and Aggregate Supply

A central bank will be concerned about the exchange rate for three reasons: (1) Movements in the exchange rate will affect the quantity of aggregate demand in an economy; (2) frequent substantial fluctuations in the exchange rate can disrupt international trade and cause problems in a nation’s banking system; (3) the exchange rate may contribute to an unsustainable balance of trade and large inflows of international financial capital, which can set the economy up for a deep recession if international investors decide to move their money to another country. Let’s discuss these scenarios in turn.

Foreign trade in goods and services typically involves incurring the costs of production in one currency while receiving revenues from sales in another currency. As a result, movements in exchange rates can have a powerful effect on incentives to export and import, and thus on aggregate demand in the economy as a whole.

For example, in 1999, when the euro first became a currency, its value measured in U.S. currency was $1.06/euro. By the end of 2013, the euro had risen (and the U.S. dollar had correspondingly weakened) to $1.37/euro. Consider the situation of a French firm that each year incurs €10 million in costs, and sells its products in the United States for $10 million. In 1999, when this firm converted $10 million back to euros at the exchange rate of $1.06/euro (that is, $10 million × [€1/$1.06]), it received €9.4 million, and suffered a loss. In 2013, when this same firm converted $10 million back to euros at the exchange rate of $1.37/euro (that is, $10 million × [€1 euro/$1.37]), it received approximately €7.3 million and an even larger loss. This example shows how a stronger euro discourages exports by the French firm, because it makes the costs of production in the domestic currency higher relative to the sales revenues earned in another country. From the point of view of the U.S. economy, the example also shows how a weaker U.S. dollar encourages exports.

Since an increase in exports results in more dollars flowing into the economy, and an increase in imports means more dollars are flowing out, it is easy to conclude that exports are “good” for the economy and imports are “bad,” but this overlooks the role of exchange rates. If an American consumer buys a Japanese car for $20,000 instead of an American car for $30,000, it may be tempting to argue that the American economy has lost out. However, the Japanese company will have to convert those dollars to yen to pay its workers and operate its factories. Whoever buys those dollars will have to use them to purchase American goods and services, so the money comes right back into the American economy. At the same time, the consumer saves money by buying a less expensive import, and can use the extra money for other purposes.

Watch It
Fluctuations in Exchange Rates

Exchange rates can fluctuate a great deal in the short run. From February 2008 to March 2009, the Indian rupee moved from 39 rupees/dollar to 51 rupees/dollar, a decline of more than one-fourth in the value of the rupee on foreign exchange markets. We read earlier that even two economically developed neighboring economies like the United States and Canada can see significant movements in exchange rates over a few years. For firms that depend on export sales, or firms that rely on imported inputs to production, or even purely domestic firms that compete with firms tied into international trade—which in many countries adds up to half or more of a nation’s GDP—sharp movements in exchange rates can lead to dramatic changes in profits and losses. So, a central bank may desire to keep exchange rates from moving too much as part of providing a stable business climate, where firms can focus on productivity and innovation, not on reacting to exchange rate fluctuations.

One of the most economically destructive effects of exchange rate fluctuations can happen through the banking system. Most international loans are measured in a few large currencies, like U.S. dollars, European euros, and Japanese yen. In countries that do not use these currencies, banks often borrow funds in the currencies of other countries, like U.S. dollars, but then lend in their own domestic currency. The left-hand chain of events in Figure 1 shows how this pattern of international borrowing can work. A bank in Thailand borrows one million U.S. dollars. Then the bank converts the dollars to its domestic currency—in the case of Thailand, the currency is the baht—at a rate of 40 baht/dollar. The bank then lends the baht to a firm in Thailand. The business repays the loan in baht, and the bank converts it back to U.S. dollars to pay off its original U.S. dollar loan.
This process of borrowing in a foreign currency and lending in a domestic currency can work just fine, as long as the exchange rate does not shift. In the scenario outlined, if the dollar strengthens and the baht weakens, a problem arises. The right-hand chain of events in Figure 1 illustrates what happens when the baht unexpectedly weakens from 40 baht/dollar to 50 baht/dollar. The Thai firm still repays the loan in full to the bank. But because of the shift in the exchange rate, the bank cannot repay its loan in U.S. dollars. (Of course, if the exchange rate had changed in the other direction, making the Thai currency stronger, the bank could have realized an unexpectedly large profit.)

In 1997–1998, countries across eastern Asia, like Thailand, Korea, Malaysia, and Indonesia, experienced a sharp depreciation of their currencies, in some cases 50% or more. These countries had been experiencing substantial inflows of foreign investment capital, with bank lending increasing by 20% to 30% per year through the mid-1990s. When their exchange rates depreciated, the banking systems in these countries became bankrupt. Argentina experienced a similar chain of events in 2002. When the Argentine peso depreciated, Argentina’s banks found themselves unable to pay back what they had borrowed in U.S. dollars.

Banks play a vital role in any economy in facilitating transactions and in making loans to firms and consumers. When most of a country’s largest banks become bankrupt simultaneously, a sharp decline in aggregate demand and a deep recession results. Since the main responsibilities of a central bank are to control the money supply and to ensure that the banking system is stable, a central bank must be concerned about whether large and unexpected exchange rate depreciation will drive most of the country’s existing banks into bankruptcy.

Try It
EXCHANGE-RATE POLICIES

Learning Objectives

- Differentiate among a floating exchange rate, a soft peg, a hard peg, and a merged currency
- Identify the tradeoffs that come with a floating exchange rate, a soft peg, a hard peg, and a merged currency

Exchange-Rate Policies

Because changes in exchange rates have macroeconomic effects on a nation’s economy, nations need to think about what exchange rate policy they should adopt. Exchange rate policies come in a range of different forms listed in Figure 1: let the foreign exchange market determine the exchange rate; let the market set the value of the exchange rate most of the time, but have the central bank sometimes intervene to prevent fluctuations that seem too large; have the central bank guarantee a specific exchange rate; or share a currency with other countries. Let’s discuss each type of exchange rate policy and its tradeoffs.

Floating Exchange Rates

A policy which allows the foreign exchange market to set exchange rates is referred to as a floating exchange rate. The U.S. dollar is a floating exchange rate, as are the currencies of about 40% of the countries in the world economy. The major concern with this policy is that exchange rates can move a great deal in a short time.

Consider the U.S. exchange rate expressed in terms of another fairly stable currency, the Japanese yen, as Figure 2 shows. On January 1, 2002, the exchange rate was 133 yen/dollar. On January 1, 2005, it was 103 yen/dollar. On June 1, 2007, it was 122 yen/dollar, on January 1, 2012, it was 77 yen per dollar, and on March 1, 2015, it was 120 yen.
per dollar. As investor sentiment swings back and forth, driving exchange rates up and down, exporters, importers, and banks involved in international lending are all affected. At worst, large movements in exchange rates can drive companies into bankruptcy or trigger a nationwide banking collapse. However, even in the moderate case of the yen/dollar exchange rate, these movements of roughly 30 percent back and forth impose stress on both economies as firms must alter their export and import plans to take the new exchange rates into account. Especially in smaller countries where international trade is a relatively large share of GDP, exchange rate movements can rattle their economies.

Figure 2. U.S. Dollar Exchange Rate in Japanese Yen. Even seemingly stable exchange rates such as the Japanese Yen to the U.S. Dollar can vary when closely examined over time. This figure shows a relatively stable rate between 2011 and 2013. In 2013, there was a drastic depreciation of the Yen (relative to the U.S. Dollar) by about 14% and again at the end of the year in 2014 also by about 14%. (Source: Federal Reserve Economic Data (FRED) https://research.stlouisfed.org/fred2/series/DEXJPUS)

However, movements of floating exchange rates have advantages, too. After all, prices of goods and services rise and fall throughout a market economy, as demand and supply shift. If an economy experiences strong inflows or outflows of international financial capital, or has relatively high inflation, or if it experiences strong productivity growth so that purchasing power changes relative to other economies, then it makes economic sense for the exchange rate to shift as well.

Floating exchange rate advocates often argue that if government policies were more predictable and stable, then inflation rates and interest rates would be more predictable and stable. Exchange rates would bounce around less, too. The Nobel prize winning economist Milton Friedman (1912–2006), for example, wrote a defense of floating exchange rates in 1962 in his book *Capitalism and Freedom*:

> Being in favor of floating exchange rates does not mean being in favor of unstable exchange rates. When we support a free price system [for goods and services] at home, this does not imply that we favor a system in which prices fluctuate wildly up and down. What we want is a system in which prices are free to fluctuate but in which the forces determining them are sufficiently stable so that in fact prices move within moderate ranges. This is equally true in a system of floating exchange rates. The ultimate objective is a world in which exchange rates, while free to vary, are, in fact, highly stable because basic economic policies and conditions are stable.

Advocates of floating exchange rates admit that, yes, exchange rates may sometimes fluctuate. They point out, however, that if a central bank focuses on preventing either high inflation or deep recession, with low and reasonably steady interest rates, then exchange rates will have less reason to vary.

**Using Soft Pegs and Hard Pegs**

When a government intervenes in the foreign exchange market so that the exchange rate of its currency is different from what the market would have produced, it is said to have established a “peg” for its currency. A soft peg is the name for an exchange rate policy where the government usually allows the exchange rate to be set by the market, but in some
cases, especially if the exchange rate seems to be moving rapidly in one direction, the central bank will intervene in the market. With a **hard peg** exchange rate policy, the central bank sets a fixed and unchanging value for the exchange rate. A central bank can implement soft peg and hard peg policies.

Suppose the market exchange rate for the Brazilian currency, the real, would be 35 cents/real with a daily quantity of 15 billion real traded in the market, as shown at the equilibrium $E_0$ in Figure 1(a) and Figure 1(b). However, the government of Brazil decides that the exchange rate should be 30 cents/real, as shown in Figure 1(a). Perhaps Brazil sets this lower exchange rate to benefit its export industries. Perhaps it is an attempt to stimulate aggregate demand by stimulating exports. Perhaps Brazil believes that the current market exchange rate is higher than the long-term purchasing power parity value of the real, so it is minimizing fluctuations in the real by keeping it at this lower rate. Perhaps the target exchange rate was set sometime in the past, and is now being maintained for the sake of stability. Whatever the reason, if Brazil’s central bank wishes to keep the exchange rate below the market level, it must face the reality that at this weaker exchange rate of 30 cents/real, the quantity demanded of its currency at 17 billion reals is greater than the quantity supplied of 13 billion reals in the foreign exchange market.

![Figure 3. Pegging an Exchange Rate.](image)

(a) Pegging an exchange rate below equilibrium  
(b) Pegging an exchange rate above equilibrium

The Brazilian central bank could weaken its exchange rate in two ways. One approach is to use an expansionary monetary policy that leads to lower interest rates. In foreign exchange markets, the lower interest rates will reduce demand and increase supply of the real and lead to depreciation. This technique is not often used because lowering interest rates to weaken the currency may be in conflict with the country’s monetary policy goals. Alternatively, Brazil’s central bank could trade directly in the foreign exchange market. The central bank can expand the money supply by creating reals, use the reals to purchase foreign currencies, and avoid selling any of its own currency. In this way, it can fill the gap between quantity demanded and quantity supplied of its currency.

Figure 3(b) shows the opposite situation. Here, the Brazilian government desires a stronger exchange rate of 40 cents/real than the market rate of 35 cents/real. Perhaps Brazil desires the stronger currency to reduce aggregate demand and to fight inflation, or perhaps Brazil believes that that current market exchange rate is temporarily lower than the long-term rate. Whatever the reason, at the higher desired exchange rate, the quantity supplied of 16 billion reals exceeds the quantity demanded of 14 billion reals.

Brazil’s central bank can use a contractionary monetary policy to raise interest rates, which will increase demand and reduce supply of the currency on foreign exchange markets, and lead to an appreciation. Alternatively, Brazil’s central bank can trade directly in the foreign exchange market. In this case, with an excess supply of its own currency in foreign exchange markets, the central bank must use reserves of foreign currency, like U.S. dollars, to demand its own currency and thus cause an appreciation of its exchange rate.

Both a soft peg and a hard peg policy require that the central bank intervene in the foreign exchange market. However, a hard peg policy attempts to preserve a fixed exchange rate at all times. A soft peg policy typically allows the exchange rate to move up and down by relatively small amounts in the short run of several months or a year, and to move by larger amounts over time, but seeks to avoid extreme short-term fluctuations.
Tradeoffs of Soft Pegs and Hard Pegs

When a country decides to alter the market exchange rate, it faces a number of tradeoffs. If it uses monetary policy to alter the exchange rate, it then cannot at the same time use monetary policy to address issues of inflation or recession. If it uses direct purchases and sales of foreign currencies in exchange rates, then it must face the issue of how it will handle its reserves of foreign currency. Finally, a pegged exchange rate can even create additional movements of the exchange rate; for example, even the possibility of government intervention in exchange rate markets will lead to rumors about whether and when the government will intervene, and dealers in the foreign exchange market will react to those rumors. Let's consider these issues in turn.

One concern with pegged exchange rate policies is that they imply a country’s monetary policy is no longer focused on controlling inflation or shortening recessions, but now must also take the exchange rate into account. For example, when a country pegs its exchange rate, it will sometimes face economic situations where it would like to have an expansionary monetary policy to fight recession—but it cannot do so because that policy would depreciate its exchange rate and break its hard peg. With a soft peg exchange rate policy, the central bank can sometimes ignore the exchange rate and focus on domestic inflation or recession—but in other cases the central bank may ignore inflation or recession and instead focus on its soft peg exchange rate. With a hard peg policy, domestic monetary policy is effectively no longer determined by domestic inflation or unemployment, but only by what monetary policy is needed to keep the exchange rate at the hard peg.

Another issue arises when a central bank intervenes directly in the exchange rate market. If a central bank ends up in a situation where it is perpetually creating and selling its own currency on foreign exchange markets, it will be buying the currency of other countries, like U.S. dollars or euros, to hold as reserves. Holding large reserves of other currencies has an opportunity cost, and central banks will not wish to boost such reserves without limit.

In addition, a central bank that causes a large increase in the supply of money is also risking an inflationary surge in aggregate demand. Conversely, when a central bank wishes to buy its own currency, it can do so by using its reserves of international currency like the U.S. dollar or the euro. But if the central bank runs out of such reserves, it can no longer use this method to strengthen its currency. Thus, buying foreign currencies in exchange rate markets can be expensive and inflationary, while selling foreign currencies can work only until a central bank runs out of reserves.

Yet another issue is that when a government pegs its exchange rate, it may unintentionally create another reason for additional fluctuation. With a soft peg policy, foreign exchange dealers and international investors react to every rumor about how or when the central bank is likely to intervene to influence the exchange rate, and as they react to rumors the exchange rate will shift up and down. Thus, even though the goal of a soft peg policy is to reduce short-term fluctuations of the exchange rate, the existence of the policy—when anticipated in the foreign exchange market—may sometimes increase short-term fluctuations as international investors try to anticipate how and when the central bank will act. The following section discusses the effects of international capital flows—capital that flows across national boundaries as either portfolio investment or direct investment.

HOW DO TOBIN TAXES CONTROL THE FLOW OF CAPITAL?

Some countries like Chile and Malaysia have sought to reduce movements in exchange rates by limiting inflows and outflows of international financial capital. This policy can be enacted either through targeted taxes or by regulations. Taxes on international capital flows are sometimes known as Tobin taxes, named after James Tobin, the 1981 Nobel laureate in economics who proposed such a tax in a 1972 lecture. For example, a government might tax all foreign exchange transactions, or attempt to tax short-term portfolio investment while exempting long-term foreign direct investment. Countries can also use regulation to forbid certain kinds of foreign investment in the first place or to make it difficult for international financial investors to withdraw their funds from a country.

The goal of such policies is to reduce international capital flows, especially short-term portfolio flows, in the hope that doing so will reduce the chance of large movements in exchange rates that can bring macroeconomic disaster.
But proposals to limit international financial flows have severe practical difficulties. Taxes are imposed by national governments, not international ones. If one government imposes a Tobin tax on exchange rate transactions carried out within its territory, the exchange rate market might easily be operated by a firm based someplace like the Grand Caymans, an island nation in the Caribbean well-known for allowing some financial wheeling and dealing. In an interconnected global economy, if goods and services are allowed to flow across national borders, then payments need to flow across borders, too. It is very difficult—in fact close to impossible—for a nation to allow only the flows of payments that relate to goods and services, while clamping down or taxing other flows of financial capital. If a nation participates in international trade, it must also participate in international capital movements.

Finally, countries all over the world, especially low-income countries, are crying out for foreign investment to help develop their economies. Policies that discourage international financial investment may prevent some possible harm, but they rule out potentially substantial economic benefits as well.

A hard peg exchange rate policy will not allow short-term fluctuations in the exchange rate. If the government first announces a hard peg and then later changes its mind—perhaps the government becomes unwilling to keep interest rates high or to hold high levels of foreign exchange reserves—then the result of abandoning a hard peg could be a dramatic shift in the exchange rate.

In the mid-2000s, about one-third of the countries in the world used a soft peg approach and about one-quarter used a hard peg approach. The general trend in the 1990s was to shift away from a soft peg approach in favor of either floating rates or a hard peg. The concern is that a successful soft peg policy may, for a time, lead to very little variation in exchange rates, so that firms and banks in the economy begin to act as if a hard peg exists. When the exchange rate does move, the effects are especially painful because firms and banks have not planned and hedged against a possible change. Thus, the argument went, it is better either to be clear that the exchange rate is always flexible, or that it is fixed, but choosing an in-between soft peg option may end up being worst of all.

Watch It

This video contrasts floating and fixed exchange rates and gives examples of ways the government could interfere in the market to affect exchange rates.

Watch this video online: https://youtu.be/_pL_5trI6YY

A Merged Currency

A final approach to exchange rate policy is for a nation to choose a common currency shared with one or more nations is also called a merged currency. A merged currency approach eliminates foreign exchange risk altogether. Just as no one worries about exchange rate movements when buying and selling between New York and California, Europeans know that the value of the euro will be the same in Germany and France and other European nations that have adopted the euro.

However, a merged currency also poses problems. Like a hard peg, a merged currency means that a nation has given up altogether on domestic monetary policy, and instead has put its interest rate policies in other hands. When Ecuador uses the U.S. dollar as its currency, it has no voice in whether the Federal Reserve raises or lowers interest rates. The European Central Bank that determines monetary policy for the euro has representatives from all the euro nations. However, from the standpoint of, say, Portugal, there will be times when the decisions of the European Central Bank about monetary policy do not match the decisions that would have been made by a Portuguese central bank.

The lines between these four different exchange rate policies can blend into each other. For example, a soft peg exchange rate policy in which the government almost never acts to intervene in the exchange rate market will look a great deal like a floating exchange rate. Conversely, a soft peg policy in which the government intervenes often to keep the exchange rate near a specific level will look a lot like a hard peg. A decision to merge currencies with another country is, in effect, a decision to have a permanently fixed exchange rate with those countries, which is like a very hard exchange rate peg. The range of exchange rates policy choices, with their advantages and disadvantages, are summarized in Table 1.

Table 1. Tradeoffs of Exchange Rate Policies
### Situation

<table>
<thead>
<tr>
<th>Floating Exchange Rates</th>
<th>Soft Peg</th>
<th>Hard Peg</th>
<th>Merged Currency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Large short-run fluctuations in exchange rates?</td>
<td>Often a lot in the short term</td>
<td>Maybe less in the short run, but still large changes over time</td>
<td>None, unless a change in the fixed rate</td>
</tr>
<tr>
<td>Large long-term fluctuations in exchange rates?</td>
<td>Can often happen</td>
<td>Can often happen</td>
<td>Cannot happen unless hard peg changes, in which case substantial volatility can occur</td>
</tr>
<tr>
<td>Power of central bank to conduct countercyclical monetary policy?</td>
<td>Flexible exchange rates make monetary policy stronger</td>
<td>Some power, although conflicts may arise between exchange rate policy and countercyclical policy</td>
<td>Very little; central bank must keep exchange rate fixed</td>
</tr>
<tr>
<td>Costs of holding foreign exchange reserves?</td>
<td>Do not need to hold reserves</td>
<td>Hold moderate reserves that rise and fall over time</td>
<td>Hold large reserves</td>
</tr>
<tr>
<td>Risk of being stuck with an exchange rate that causes a large trade imbalance and very high inflows or outflows of financial capital?</td>
<td>Adjusts often</td>
<td>Adjusts over the medium term, if not the short term</td>
<td>May become stuck over time either far above or below the market level</td>
</tr>
</tbody>
</table>

Global macroeconomics would be easier if the whole world had one currency and one central bank. The exchange rates between different currencies complicate the picture. If exchange rates are set solely by financial markets, they fluctuate substantially as short-term portfolio investors try to anticipate tomorrow's news. If the government attempts to intervene in exchange rate markets through soft pegs or hard pegs, it gives up at least some of the power to use monetary policy to focus on domestic inflations and recessions, and it risks causing even greater fluctuations in foreign exchange markets.

There is no consensus among economists about which exchange rate policies are best: floating, soft peg, hard peg, or merged currencies. The choice depends both on how well a nation's central bank can implement a specific exchange rate policy and on how well a nation's firms and banks can adapt to different exchange rate policies. A national economy that does a fairly good job at achieving the four main economic goals of growth, low inflation, low unemployment, and a sustainable balance of trade will probably do just fine most of the time with any exchange rate policy; conversely, no exchange rate policy is likely to save an economy that consistently fails at achieving these goals. On the other hand, a merged currency applied across wide geographic and cultural areas carries with it its own set of problems, such as the ability for countries to conduct their own independent monetary policies.
Learning Objectives

**floating exchange rate**: a country lets the exchange rate market determine its currency’s value

**hard peg**: an exchange rate policy in which the central bank sets a fixed and unchanging value for the exchange rate

**international capital flows**: flow of financial capital across national boundaries either as portfolio investment or direct investment

**merged currency**: when a nation chooses to use another nation’s currency

**soft peg**: an exchange rate policy in which the government usually allows the market to set the exchange rate, but in some cases, especially if the exchange rate seems to be moving rapidly in one direction, the central bank will intervene

**Tobin taxes**: see international capital flows

---

**PUTTING IT TOGETHER: EXCHANGE RATES AND INTERNATIONAL FINANCE**

The goal of this module was to explain how a nation’s currency exchange rate affects its balance of trade and the state of the macroeconomy.
You learned about how:

- The foreign exchange market works
- Importers and exporters affect the demand and supply of foreign exchange
- Portfolio and direct investment affect the demand and supply of foreign exchange
- Changes in exchange rates cause changes in the trade balance
- How changes in the trade balance (surplus or deficit) affect the domestic economy
- How changes in exchange rates influence a nation’s financial system
- To differentiate between different currency

You learned that since exchange rate changes can affect the both product markets and financial markets, nations need to carefully choose their exchange rate policies, since making the wrong choice can have unforeseen impacts on the economy. You also learned about appreciating and depreciating currencies, and that there are pros and cons to each.

Do you think it’s best for the United States to have a stronger currency?

**IS A STRONGER CURRENCY GOOD FOR ECONOMY?**

The foreign exchange value of the dollar is a price and whether a higher price is good or bad depends on where you are standing: sellers benefit from higher prices and buyers are harmed. A stronger dollar is good for U.S. imports (and people working for U.S. importers) and U.S. investment abroad. It is also good for U.S. tourists going to other countries, since their dollar goes further. But a stronger dollar is bad for U.S. exports (and people working in U.S. export industries); it is bad for foreign investment in the United States (leading, for example, to higher U.S. interest rates); and it is bad for foreign tourists (as well as U.S hotels, restaurants, and others in the tourist industry). In short, whether the U.S. dollar is good or bad is a more complex question than you may have thought. The economic answer is “it depends.” What we can say is that a stable U.S. dollar is good for the economy so that people can make their purchases and businesses can produce and sell without worrying about the value of the dollar in foreign exchange markets.